2000

Novel analysis, decomposition and reconstruction techniques for waveform interpolation speech coding

Nicola Raewyn Chong-White
University of Wollongong

Novel Analysis, Decomposition and Reconstruction Techniques for Waveform Interpolation Speech Coding

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

by

Nicola Raewyn Chong-White

Bachelor of Engineering (Electrical and Electronic) (Honours I)

School of Electrical, Computer and Telecommunications Engineering

July 2000

Abstract

Speech coding has experienced rapid growth throughout the past decade as many new desirable commercial applications have emerged. However, there remains a void of solutions for good synthesised speech at bit rates around 4kbit/s. At this transmission rate, the limitations of both waveform coders, which produce excellent quality at higher rates, and parametric coders, which operate well at lower rates, inhibit their performance quality.

In this thesis, several techniques that provide improved signal analysis and bridge the gap between waveform and parametric coders, are proposed. Firstly, basic Waveform Interpolation (WI) principles are considered. These take advantage of the pitch periodicity and perceptual redundancies of speech. The decomposition of WI, which separates voiced and unvoiced characteristics, is an advantageous mechanism to exploit perceptual differences. This concept is thus extended to provide a multi-resolution analysis of speech evolution by implementing perfect reconstruction wavelet filter banks. Several causal, stable, finite impulse response (FIR) and infinite impulse response (IIR) filter bank designs are discussed. These are adapted to the signal properties by drawing upon closely related wavelet theory. The proposed wavelet decomposition allows the application of flexible, efficient, perception-based quantisation techniques to code the excitation signal.

To achieve toll-quality at 4kbit/s, this thesis postulates that it is advantageous to integrate the waveform coding property into the WI paradigm. Perfect reconstruction WI coders have been previously described and these rely on accurate time-warping to operate effectively. However, a technique to achieve this was not established. The proposed Waveform-Matched WI (WMWI) coder aims to address

this issue, providing an improved means of signal analysis which enables efficient quantisation and achieves waveform matching. Several analysis techniques are developed, including the formation of an optimal pitch contour to ensure the warped signal has an exactly constant pitch, and various decomposition mechanisms to provide for efficient quantisation. The techniques contrast conventional WI methods by preserving all input samples and facilitating waveform coding.

The WMWI reconstruction achieves waveform matching by directly inverting the analysis procedures. A novel pitch quantisation technique is presented, in which important facets of the pitch track are identified and transmitted. An accurate representation of the pitch track can then be formed, enabling time-synchrony between the input and synthesised speech. The preferred methods for effectively quantising the decomposed components of the excitation signal and taking full advantage of the warped-domain representation, are also discussed. The quantisation of the slowly-evolving component, which contains the periodic structure of speech, derives clear benefit from the warping process, and is therefore performed in the warped time domain. In particular, the warping facilitates high performance waveform matching through VQ techniques. Conversely, the rapidly-evolving component displays no periodicity. Hence, the advantages of a warped domain representation for this component lie in the ability to perform fast, constant length, transforms on consecutive pitch cycles.

At 4kbit/s, the WMWI coder produces output speech with quality exceeding that of both the standardised 2.4kbit/s MELP coder and 4.8kbit/s FS-1016 CELP coder. In addition, the perfect reconstruction property of the WMWI paradigm and its advanced methods which ensure time-synchrony, facilitate improved quality speech coding at higher bit rates over exclusively parametric approaches.

Contents

Statement of Originality

I, Nicola R. Chong-White, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, is wholly my own work, except where due reference is made in the text.

This document has not been submitted for qualifications at any other academic institution.

Signed

Nicola R. Chong-White

7 July 2000

Acknowledgments

Firstly, I would like to thank my supervisor, Dr. Ian Burnett, for his continuous guidance and support, and for encouraging me to pursue all opportunities. Also thanks to Professor Joe Chicharo and Dr. Mark Thomson for their helpful suggestions and discussions throughout the past few years.

I would also like to thank my colleagues, both past and present, in the Whisper Laboratories. The atmosphere created in the lab has been wonderful to work in, and I hold with me some unforgettable memories.

To my parents, I would like to express my gratitude for supporting my decision to study abroad and for giving me the strength to chase my ambitions.

Finally, my most heartfelt thanks go to my husband, Christian, who commuted for endless hours a day so that I could focus on my research. His love and support has been truly invaluable.

List of Figures

List of Tables

List of Abbreviations

ADPCM	Adaptive Differential Pulse Code Modulation
CCITT	International Consultative Committee for Telephone and Telegraph
CELP	Code-Excited Linear Prediction
CW	Characteristic Waveform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
ELT	Extended Lapped Transform
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GSM	Global System for Mobile Communications
GT	Gabor Transform
IIR	Infinite Impulse Response
IMBE	Improved Multi-band Excitation Coder
ITU-T	International Telecommunications Union – Telecommunications Standardisation Sector
LOT	Lapped Orthogonal Transform
LP	Linear Prediction
LPAS	Linear Prediction Analysis-by-Synthesis
LPC	Linear Predictive Coding

LSF	Line Spectral Frequency
LTP	Long-Term Prediction
MBE	Multi-band Excitation
MELP	Mixed Excitation Linear Prediction
MLT	Modulated Lapped Transform
MPE	Multi-Pulse Excitation
MSE	Mean Squared Error
PCM	Pulse Code Modulation
PSWT	Pitch-Synchronous Wavelet Transform
PWI	Prototype Waveform Interpolation
QMF	Quadrature Mirror Filter Bank
RCELP	Relaxed Code-Excited Linear Prediction
REW	Rapidly Evolving Waveform
RPE	Regular Pulse Excitation
SEW	Slowly Evolving Waveform
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
VDVQ	Variable Dimension Vector Quantisation
VQ	Vector Quantisation
WI	Waveform Interpolation
WMWI	Waveform-Matched Waveform Interpolation
WT	Wavelet Transform

Chapter 1

Introduction

You see, wire telegraph is a kind of a very, very long cat.
You pull his tail in New York and his head is meowing in Los Angeles.
Do you understand this? And radio operates exactly the same way:
you send signals here, they receive them there.
The only difference is that there is no cat."

*— **Albert Einstein***

1.1 Speech Coding at Low Bit-Rates

Speech coding, or speech compression, has advanced rapidly over the past decade with the emergence of many appealing commercial applications. Wireless cellular communications, in particular, have become increasingly popular, and the growth of the internet has led to a corresponding growth of multimedia communications, requiring simultaneous transmission of voice and data. Other applications include satellite communications, visual telephony, and message retrieval/storage systems. The improvements in computing technology have allowed the efficient implementation of complex speech coding algorithms integral to these applications.

The drive for further speech coding research is motivated by the need to save bandwidth across telecommunications networks, which are becoming saturated by the escalation in user activity and a diversifying user base. To obtain good performance at low rates, speech coders of current interest exploit characteristics of speech production and perception. This enables redundancies to be removed from the signal, leaving only perceptually significant information to be transmitted. As research proceeds to obtain a better understanding of the perceptual nature of speech, the bit rate required to achieve toll-quality will continue to decline (though the reductions achievable within reasonable complexity requirements are diminishing in size).

In this thesis, several methods for efficient coding of a speech signal are examined. The methods can be divided into four stages:

i) *analysis techniques,* whereby redundancies in the input speech signal are identified and exploited,

ii) *decomposition techniques,* allowing signal components with distinct properties to be separated for increased coding efficiency,

iii) *quantisation techniques,* providing an efficient representation for each parameter allowing effective transmission across the channel, and

iv) *reconstruction techniques,* whereby the transmitted parameters are combined to recreate a perceptually accurate reproduction of the original signal at the decoder.

These methods are implemented within the Waveform Interpolation (WI) framework, although many of the findings in this thesis, with regard to quantisation techniques, could equally be applied to other coding architectures.

1.2 Outline of this Thesis

The organisation of this thesis is as follows: Chapter 2 reviews the development of speech coding techniques to improve the quality of quantised speech, while allowing transmission at low data rates. Successful methods include the use of linear and pitch prediction to remove signal correlations, perceptual weighting and postfiltering to adapt the power spectrum of the speech to be more pleasing to the human ear, analysis-by-synthesis to provide feedback in the choice of parameters, and decomposition techniques to separate the signal into components which are more convenient for quantisation. A detailed overview of the Waveform Interpolation (WI) coder, and its initial modification from a parametric coder to a waveform coder, is also presented.

In Chapter 3, an alternative decomposition mechanism is proposed for the WI coder. The decomposition is based on an evolutionary domain wavelet transform, which is adapted to real-time speech coding applications. The wavelet decomposition (implemented as perfect reconstruction filter banks) separates the evolution of pitch-length segments of speech into several frequency subbands of decreasing resolution. This allows effective perception-based quantisation techniques to be applied to a series of reduced resolution component waveforms. Several types of filter banks are investigated, and the phase inter-relationships between the decomposed surfaces

(which result in reconstruction difficulties) are discussed. This leads to a preferred decomposition/reconstruction configuration.

Chapter 4 presents the Waveform-Matched Waveform Interpolation (WMWI) technique; an adaptation of the WI paradigm to achieve perfect reconstruction of the input signal. The key to WMWI analysis is the formation of an optimised pitch track, which will enable accurate time-warping of the speech residual, such that it has an exactly constant pitch period. This facilitates the effective use of fixed-length, pitch-synchronous operations (e.g. transforms, decompositions) to produce parameters that may be efficiently quantised. The analysis techniques are completely invertible and do not destroy any speech information in the unquantised case.

The focus of Chapter 5 is the incorporation of quantisation and reconstruction techniques into the WMWI framework, which provide an accurate, time-synchronous description of the input signal. Significant emphasis is placed on the effective transmission and recreation of the pitch contour. This is central to the waveform matching objective. The quantisation of slowly-evolving and rapidly-evolving signal components is also discussed in relation to the benefits of time-warping. Comparison is made with standard WI which assumes linear interpolation of pitch and hence, does not attempt to preserve time-synchrony or waveform detail. The subjective performance results of the 4kbit/s WMWI coder are discussed in comparison to the standard coders of G.729, an 8kbit/s CELP coder, FS-1016, a 4.8kbit/s CELP coder and the 2.4kbit/s MELP Federal Standard.

Finally, Chapter 6 summarises the main findings of this thesis and identifies areas for further research.

1.3 Contributions

A list of the major contributions of this thesis is given below. They are sorted in order of appearance, with the corresponding chapter references and associated publications shown in parentheses.

- Performed an analysis of the WI coder in the presence of noise, motivating the search for alternative decomposition techniques. [Chon97]
- Introduced the Pitch Synchronous Wavelet Transform (PSWT) as a decomposition mechanism for the WI speech coding paradigm. The adaptation of the PSWT to a real-time coder requires the incorporation of the characteristic waveform (CW) extraction and alignment techniques of WI. (Chapter 3.5) [Chon98a][Chon00a]
- Developed the PSWT as a multi-scale decomposition of the evolving CW surface. The standard PSWT implementation using FIR QMF banks is inappropriate for real-time coding due to delay restrictions, thus, a stable, causal IIR QMF solution was developed. The sharper filter responses improve the capture of CW dynamics and contrast with the FIR filters of the SEW/REW approach. Low-delay FIR filter banks were also applied to achieve reduced delay. (Chapter 3.6) [Chon98a] [Chon98b] [Chon99a]
- Evaluated the perceptual significance of evolutionary frequency subbands by subjective testing. The results form the basis for the designated bit

allocations for these components. This provides for improved flexibility and scalability. (Chapter 3.9.5)

- Identified the phase inter-relationships between the subbands within a tree-structured decomposition. These relationships must be substantially maintained in order to obtain good quality reconstructed speech. (Chapter 3.9.7) [Chon99a]
- Illustrated the crucial importance of accurate warping to enable effective decomposition within perfect reconstruction WI, by studying the effects of poor alignment in the time and frequency domains. (Chapter 4.3.3) [Chon99b]
- Enhanced the reliability of perfect reconstruction WI by designing a technique to ensure perfect alignment of the pitch periods of voiced speech, even after unvoiced segments. (Chapter 4.4) [Chon99b] [Chon00c]
- Developed a technique to locate pitch pulses using the contributions of weighted autocorrelation functions at different offsets, determining the existence of a pulse based on several criteria, including adaptive thresholding. (Chapter 4.4.1) [Chon99b]
- Formulated a reliable technique to create an optimal pitch track of a speech signal, which allows the signal to be pitch normalised with the pitch pulse occurring at the central sample of the pitch period. This pitch track is defined for four possible frame types. (Chapter 4.4.4) [Chon00b]
- Identified the significant characteristics of the pitch contour which need to be preserved, to enable the formation of an accurate description of the pitch

track at the decoder. This is done by quantising one pitch value per frame, plus additional side information. (Chapter 5.3.2) [Chon00b]

- Formed a method to accurately recreate the pitch track, enabling time-synchronous reconstruction, and hence allowing waveform matching of the speech signal. This avoids the possibility of some pitch periods being repeated or omitted as in standard WI, which results in speech artefacts. (Chapter 5.4) [Chon00b][Chon00d][Chon00e]
- Showed the advantages and disadvantages of representing pitch-cycle waveforms in the warped (constant pitch) and unwarped time domains, and used these findings, in addition to perceptual knowledge, to determine the quantisation techniques for the voiced and unvoiced parts of the speech signal. (Chapters 5.6 and 5.7)
- Implemented a 4kbit/s speech coder which produces high quality output speech, exceeding that of the standardised 4.8kbit/s FS-1016 CELP and 2.4kbit/s MELP coders. (Chapter 5.9)

1.4 Publications

Below is a list of the publications arising from the research presented in this thesis.

1.4.1 Journal Publications

- N. R. Chong, I. S. Burnett and J. F. Chicharo, "A new Waveform Interpolation coding scheme based on pitch synchronous wavelet transform

decomposition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 345-348, May 2000. [Chon00a]

- N. R. Chong and I. S. Burnett, "Accurate, critically-sampled characteristic waveform surface construction for Waveform Interpolation decomposition," *IEE Electronics Letters*, vol. 36, pp. 1245-1247, 6 July 2000. [Chon00e]
- N. R. Chong and I. S. Burnett, "Improved signal analysis and waveform matching in Waveform Interpolation speech coding," submitted to *IEEE Trans. Speech and Audio Processing*, July 2000.

1.4.2 Conference Publications

- N. R. Chong, I. S. Burnett, J. F. Chicharo and M. M. Thomson, "The effects of noise on the Waveform Interpolation speech coder," *Proc. IEEE TENCON, Annual Conf. Speech, Image Technol. Comput., Telecommun.*, Brisbane, Australia, vol. 2, pp. 609-612, Dec. 1997. [Chon97]
- N. R. Chong, I. S. Burnett, J. F. Chicharo and M. M. Thomson, "Use of the pitch synchronous wavelet transform as a new decomposition method for WI," *Proc. IEEE Int Conf. Acoust., Speech, Signal Processing*, Seattle, USA, vol. 1, pp. 513-516, May 1998. [Chon98a]
- N. R. Chong, I. S. Burnett and J. F. Chicharo, "An improved decomposition method for WI using IIR filter banks," *Proc. 5th Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998. [Chon98b]
- N. R. Chong, I. S. Burnett and J. F. Chicharo, "Low delay multi-level decomposition and quantisation techniques for WI coding," *Proc. IEEE Int*

Conf. Acoust., Speech, Signal Processing, Phoenix, USA, vol. 1, pp. 241-244, Mar. 1999. [Chon99a]

- N. R. Chong, I. S. Burnett and J. F. Chicharo, "Adapting Waveform Interpolation (with pitch-spaced subbands) for quantisation", *Proc. IEEE Speech Coding Workshop*, Porvoo, Finland, pp. 96-98, Jun. 1999. [Chon99b]
- N. R. Chong and I. S. Burnett, "Improved signal analysis and time-synchronous reconstruction in Waveform Interpolation coding," *Proc. IEEE Speech Coding Workshop*, pp. 56-58, Delavan, USA, Sep. 2000. [Chon00b]

1.4.3 Patents

- N. R. Chong and I. S. Burnett, "Method and apparatus for time-warping a digitised signal to have an approximately fixed period," Australian Provisional Patent, filed Feb. 1999 (replaced by US Patent CR1029AC, filed Feb. 2000). [Chon00c]
- N. R. Chong and I. S. Burnett, "Method and apparatus for encoding and reconstructing the pitch track of a digitised time-varying waveform," Australian Provisional Patent, filed Apr. 2000. [Chon00d]

Chapter 2

Effective and Efficient Representations of Speech Signals

The outcome of any serious research can only be
to make two questions grow where only one grew before.

-- ***Thorstein Bunde Veblen***

2.1. Introduction

In recent years, with the rapid development and growth of cellular and satellite communications systems, there has become an increasingly strong need for the efficient representation of speech signals. This requires minimising the level of information needed to represent the signal to maximise channel utilisation, while still maintaining high perceptual output quality. To achieve this improved performance, coding techniques must exploit the properties of human speech production and perception.

Effective methods for reducing the bit rate include:

- Short-Term Prediction,
- Long-Term (Pitch) Prediction

- Vector Quantisation,
- Exploitation of Perceptual Indifferences,
- Signal Transformation, and
- Signal Decomposition

These are combined with methods for improving the speech quality, which may be slightly degraded due to the above rate-reduction techniques. Methods commonly used are:

- Perceptual Weighting, and
- Post-filtering

In this chapter, we review a number of speech coding algorithms, focussing on the techniques used to increase coding efficiency and improve perceptual quality.

2.2. Speech Production and Perception

To effectively code speech, knowledge of how the human vocal system produces speech sounds, as well as how the human auditory system perceives these sounds, is required. These factors contribute significantly to the design of the speech coder. A speech coder ideally removes all redundant information, producing an efficient representation of the speech signal to be transmitted.

2.2.1. Speech Production

Speech is produced when air is forced from the lungs through the vocal tract [Flan72][Clar95]. The airflow may or may not be restricted or momentarily blocked by the vocal cords, situated in the vocal tract. Speech can be broadly classified into

two types: voiced and unvoiced (Figure 2.1). Voiced sounds, e.g. vowels /a/, /i/, are produced when air from the lungs causes oscillation of the vocal cords. The resulting speech is quasi-periodic, and its fundamental frequency, or pitch, is determined by the rate at which the vocal cords vibrate. In general, male speakers produce speech with a lower fundamental frequency than female speakers. Unvoiced sounds are produced when the air flow from the lungs is forced through a constriction in the vocal tract. The nature of these sounds, for example, fricatives, e.g. /f/, /s/, plosives, e.g. /t/, /p/ or nasals, e.g. /m/, /n/, depends on the degree of the constriction. Unvoiced sounds do not exhibit any periodicity and possess a noisy character. Many parts of speech display a mixture of both voiced and unvoiced sounds.

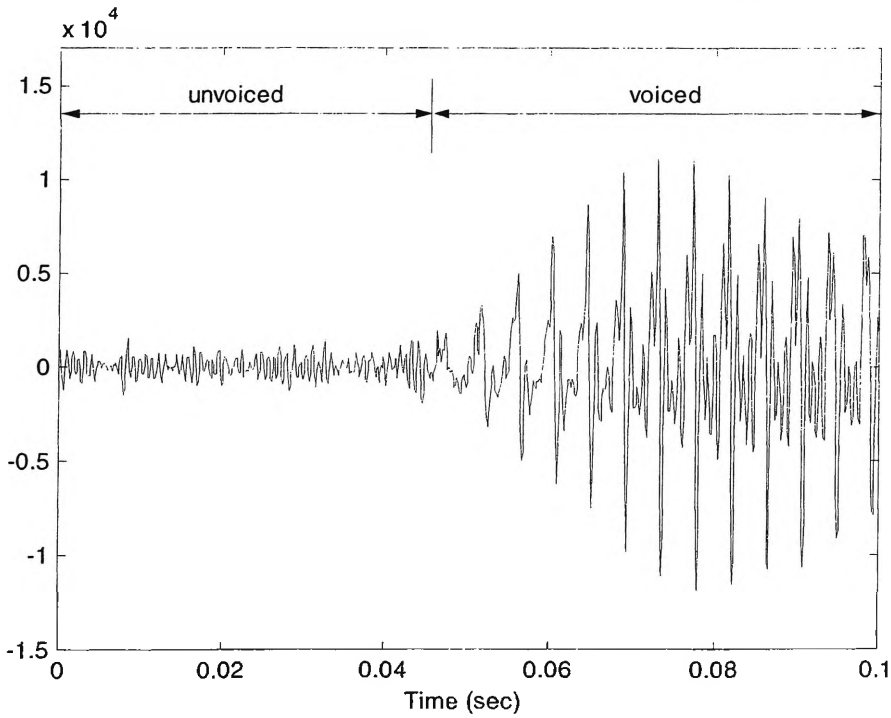


Figure 2.1. Time-domain representation of a speech signal

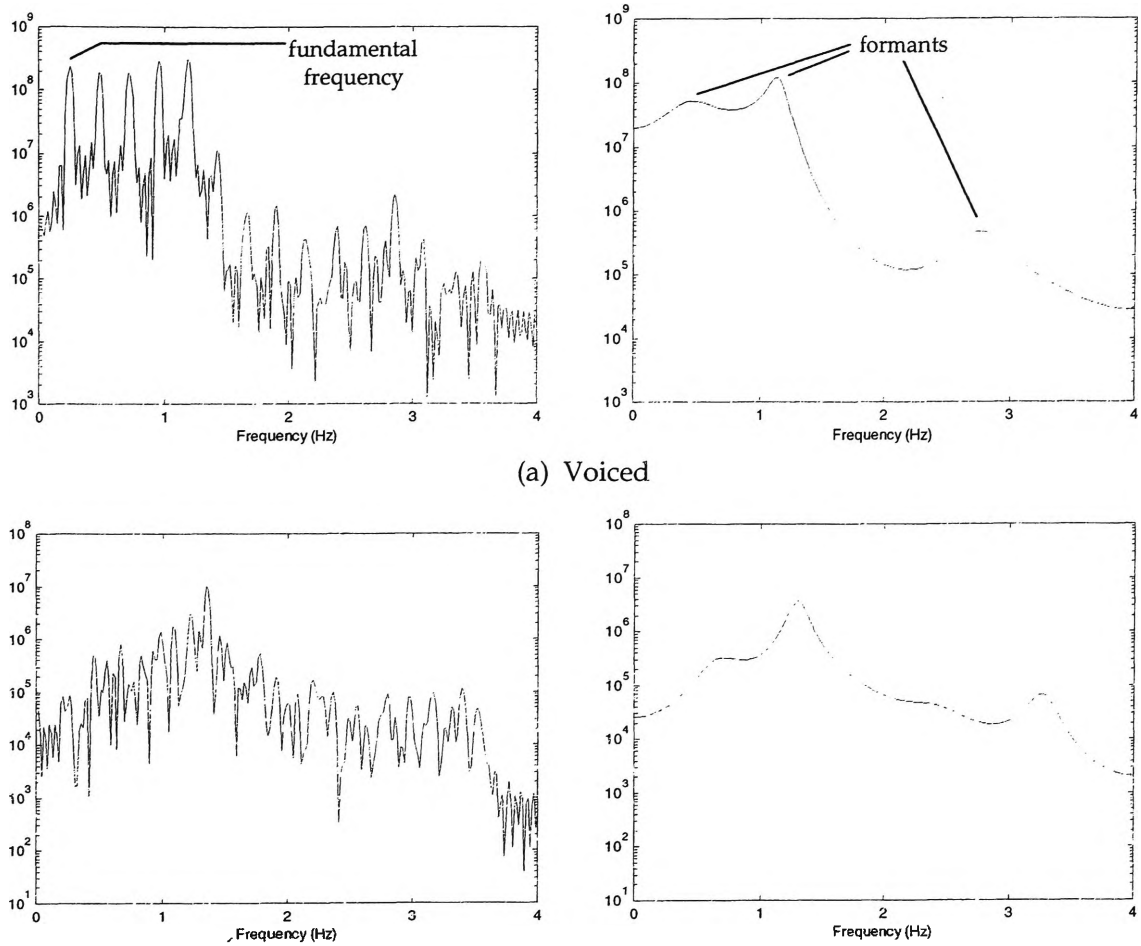


Figure 2.2. The short-time power spectrum (left) and spectral envelope of a voiced sound and unvoiced sound. (20ms window length)

Properties of a speech signal

While speech is a non-stationary signal, it can be considered to be quasi-stationary over short segments. Hence, analysis is performed on a short-time basis over segments of 20-30ms in duration. The short-term power spectrum of a voiced speech sound has two main features: the spectral envelope, due to the shape of the vocal tract, and a harmonic (fine) structure, formed by the vibrations of the vocal cords [Rabi78][Clar95]. The spectral envelope is characterised by a set of broad peaks, called formants, which are the resonant modes of the vocal tract. The

harmonic structure consists of sharp peaks, occurring at multiples of the fundamental frequency. For unvoiced sounds, this harmonic structure is not present. These characteristics are presented in Figure 2.2.

The spectral envelope corresponds to short-term correlations in the speech, and the spectral fine structure relates to the long-term correlations. These correlations imply that redundancy exists in the signal, which should be removed during coding to improve efficiency.

2.2.2. Speech Perception

The human auditory system can perceive sounds across a wide frequency range, however, not all sounds are perceived in the same manner. One non-linear phenomenon of the perception of sounds is the effect of masking [Flan72][Moor97]. When two different sounds occur simultaneously, or at close intervals to each other, one sound may "mask" the other, rendering it inaudible. A signal is most easily masked by a sound of higher energy that contains similar frequency components. This suggests masking reflects the limits on which the ear can successfully separate the two sounds. Frequency selectivity is often expressed in term of critical bands (the Bark Scale), in which frequency resolution broadens logarithmically with increasing frequency.

The human ear also has very different perception of voiced and unvoiced sounds. Since unvoiced sounds have less redundancy due to their non-periodic nature, to reach a desired SNR ratio, they require a higher bit rate than for voiced sounds.

However, a lower bit rate is sufficient to achieve the same perceptual quality. In fact, while small changes in the stimulus of vowel sounds are easily perceived, unvoiced sounds can be replaced by a different signal with noise-like fine structure and similar spectral envelope without degradation in the perceived output quality [Kubi93].

2.3. Types of Speech Coders

The objective of speech coding is to efficiently code the significant components of the input speech signal to achieve a certain level of output quality. The acceptable quality requirement depends on the intended application. Speech coders can be classified into two main categories:

- Waveform coders, and
- Parametric coders (also called vocoders).

Waveform coders, eg. Pulse Code Modulation (PCM), Adaptive Differential PCM (ADPCM) [Jaya84], are characterised by the property that the reconstructed signal converges to the original signal with decreasing quantisation error (increasing bit rate). They are very successful for the coding of speech at medium and high bit rates, however, at bit rates below 4kbit/s the speech quality degrades rapidly due to the inability to adequately approximate the original waveform.

Parametric coders, e.g. LPC-10 [Trem82], sinusoidal coders [McAu95], are based on a mathematical model of speech production and perception. They construct a new and different waveform that has a perceptually similar sound to the input signal,

but however, do not attempt to approach the original signal with increasing bit rate. Instead, the speech quality at high rates is limited by the accuracy of the speech production model. Parametric coders are popular for coding at low bit rates where effective waveform reproduction cannot be achieved.

2.4. Pulse Code Modulation

The simplest method of forming a digital representation of speech is by using Pulse Code Modulation (PCM) [Jaya84]. In PCM, an analog waveform is sampled in time, and its sample amplitudes are individually rounded to the nearest quantisation level. The quantiser may have a uniform, or non-uniform, e.g. logarithmic, step-size. While uniform PCM has very low complexity, it does not exploit any speech properties and thus, is expensive in terms of bit rate. However, efficiency can be enhanced by selecting the quantiser levels to better match the distribution of the speech samples. This is achieved using logarithmic PCM, which uses a small step-size for frequently occurring amplitudes and a larger step size for amplitudes which occur less often. It also exploits the fact that the human ear is less sensitive to small errors for sounds of high intensity. 64kbits/s log-PCM with either μ-law or A-law companding became the first speech coding algorithm standardised by the CCITT[1] (G.711), and its output quality is considered "toll-quality"; a reference against which the performance of other coding strategies is measured.

[1] International Consultative Committee for Telephone and Telegraph. It is currently known as the International Telecommunications Union – Telecommunications Standardisation Sector (ITU-T).

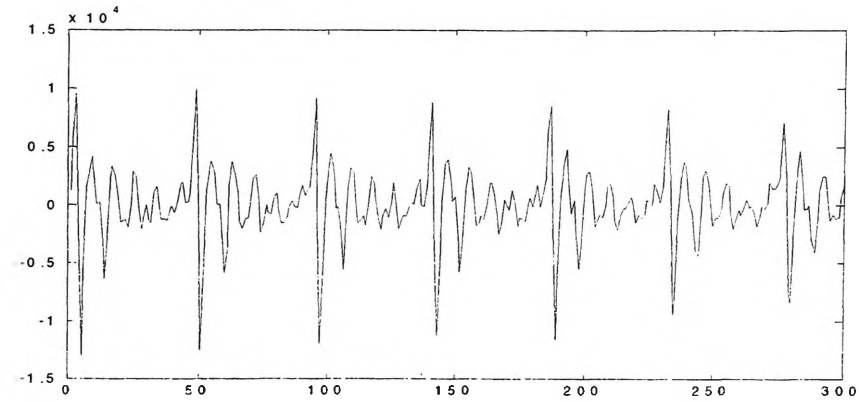
PCM, however, does not take advantage of the correlations between adjacent speech samples. Adaptations of the technique which exploit these correlations are Differential PCM (DPCM), Delta Modulation (DM), and Adaptive DPCM (ADPCM) [Gibs80]. In ADPCM, prediction is used to estimate the current speech sample from previous sample values, and the prediction error is quantised. Both the predictors and quantiser step-size adapt with the time-varying characteristics of the speech. The technique achieves toll-quality performance at 32kbits/s and was adopted for the CCITT standard G.721. This illustrates the ability to reduce the required bit rate without sacrificing quality, by adapting the quantisation technique to the statistics of the signal.

2.5. Linear Prediction

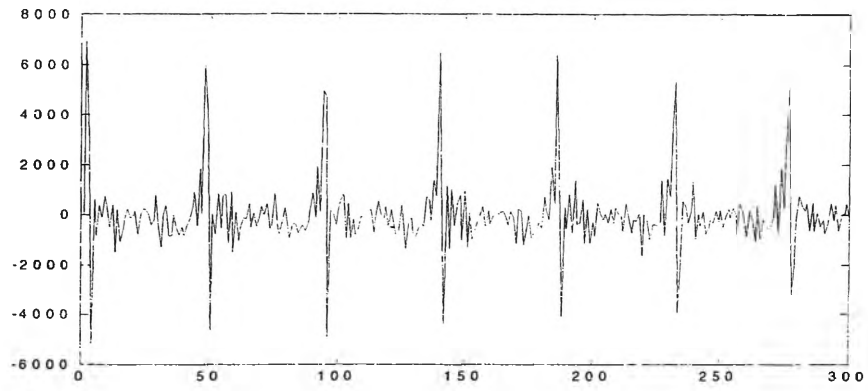
To further reduce the required speech transmission rate, it is advantageous to exploit its basic structure. The spectral envelope, formed by the vocal tract, introduces redundancy in the form of short-term correlations between speech samples. This redundancy can be removed using linear prediction (LP) techniques, whereby an autoregressive model is fitted to the spectrum of a short speech section [Makh75][Mark76]. In LP, a predictor is used to predict the current speech sample from P previous samples. This can be written as:

$$\begin{aligned}\tilde{s}(n) &= a_1 s(n-1) + a_2 s(n-2) + \ldots + a_P s(n-P) \\ &= \sum_{k=1}^{P} a_k s(n-k)\end{aligned} \tag{2.1}$$

where a_k are the linear prediction coefficients, and P is the order of prediction. The LP coefficients can be determined by LP analysis techniques such as the



(a) Speech signal



(b) Residual Signal

Figure 2.3. Diagram of a speech signal and its corresponding residual signal.

autocorrelation or covariance method [Mark76]. The coefficients are then transformed to a set of parameters, such as Line Spectral Frequencies (LSFs), which allow more efficient coding.

The resulting output, or residual signal,

$$\begin{aligned} e(n) &= s(n) - \tilde{s}(n) \\ &= s(n) - \sum_{k=1}^{P} a_k s(n-k) \end{aligned} \tag{2.2}$$

can be thought of as the output of a LP analysis filter with the transfer function

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k} . \tag{2.3}$$

The residual has a low level of redundancy between adjacent samples and a significantly flatter power spectrum, making the quantisation task more efficient. To synthesise speech, the residual may simply be passed through the inverse filter, $\frac{1}{A(z)}$. A diagram showing the input speech signal and its corresponding residual signal is in Figure 2.3.

2.6. Linear Prediction Parametric Coders

2.6.1. Linear Predictive Coding Vocoder

An early speech coder, which uses linear prediction analysis, is the Linear Predictive Coder (LPC) [Atal71][Trem82]. This is a parametric coder, based on a simplified source-system model of speech production, as shown in Figure 2.4. In the model, a binary voicing decision determines whether the generated filter excitation signal is either a periodic pulse train (to represent voiced residual) or Gaussian noise (to represent unvoiced residual). The vocal tract may be represented by a time-varying filter, which can be described by an all-pole (autoregressive) model of the form:

$$\begin{aligned} H(z) &= \frac{G}{A(z)} \\ &= \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \end{aligned} \tag{2.4}$$

where a_k are the filter coefficients, G is a gain factor and P is the filter order. The parameters of this filter are estimated using LP analysis, and provide an efficient

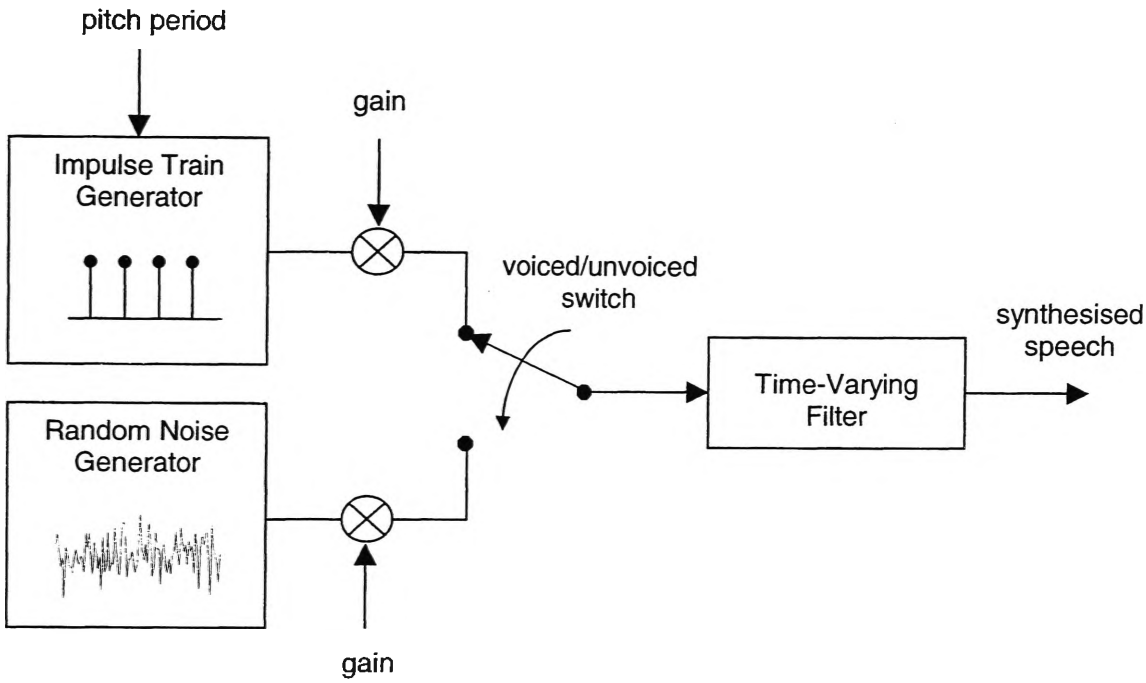


Figure 2.4. Simplified source filter model of speech production used in the LPC vocoder synthesiser

representation of the spectral envelope of the speech. This is used to shape the flat power spectrum of the excitation to that of the speech signal.

The model requires only a small amount of information to be transmitted across the channel – the voicing decision, pitch, filter coefficients and gain. No information is sent to specify the detail of the excitation waveform. The coder, known as LPC-10 due to its 10^{th} order linear prediction synthesis filter, was adopted as a government standard, FS-1015, for secure voice communication at 2.4 kbit/s [Trem82].

2.6.2. Mixed-Excitation LPC Vocoder

A disadvantage of LPC vocoders is that the reproduced speech suffers from an unnatural, buzzy quality, especially in the presence of acoustical background noise. The main source of speech degradation can be attributed to the binary voicing

decision, as transitional regions of speech cannot be modelled well, and any error in the voicing classification will cause significant distortion. A number of improvements have been made to the basic LPC vocoder by using a mixture of pulse and noise excitation [Makh78][Kang85] to reduce the synthetic speech quality. More recently, a detailed Mixed Excitation Linear Prediction (MELP) model was proposed by McCree *et al.* [McCr95], which introduces additional attributes such as periodic or aperiodic pulses, adaptive spectral enhancement and a pulse dispersion filter. These features can better represent speech characteristics and enhance performance by allowing different mixtures of pulse and noise excitation for each of a number of frequency bands. The MELP coder was selected as the new Federal Standard for 2.4kbit/s voice communications [Supp97][McCr96], and comparative tests have indicated that its quality exceeds that of the 4.8kbit/s government standard Code-Excited Linear Prediction [Kohl97]. MELP techniques are currently under consideration for 4kbit/s standardisation.

2.7. Long-Term (Pitch) Prediction

While LP analysis effectively removes the short-term correlations present in speech, long-term correlations still exist, especially during voiced regions. Hence, a second predictor, a long-term (or pitch) predictor (LTP), is cascaded with the short-term predictor to remove the periodicity of the residual [Rama87, 89]. A LTP subtracts past delayed, gain-adjusted sections of the LP residual signal which maximally correlate the current section. Figure 2.5 depicts the operation of the LTP analysis filter.

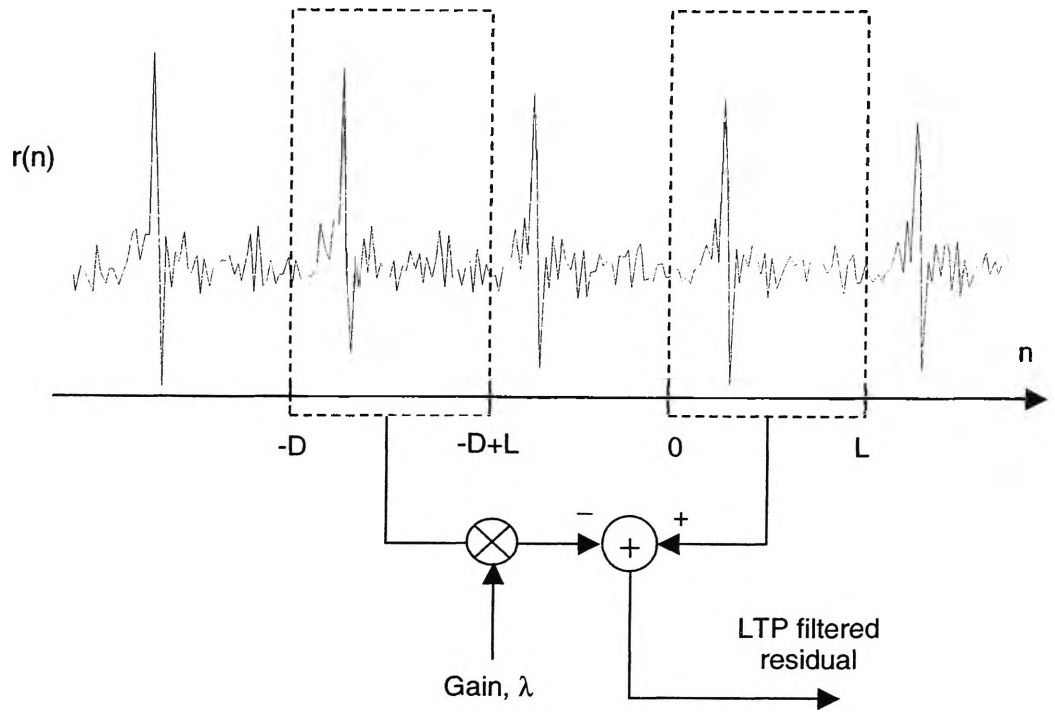


Figure 2.5. The long-term prediction analysis filtering operation

The squared error, for a segment of length L, can be expressed as:

$$e(D,\lambda)=\sum_{n=0}^{L-1}\left[r(n)-\lambda r(n-D)\right]^2 \tag{2.5}$$

where D is the delay and λ is the gain.

Thus, a solution for the optimum D and λ is determined by setting $\frac{\partial e(D,\lambda)}{\partial \lambda}=0$, giving:

$$\lambda=\frac{\sum_{n=0}^{L-1} r(n)r(n-D)}{\sum_{n=0}^{L-1} r^2(n-D)}. \tag{2.6}$$

Substituting Equation (2.6) into (2.5), leads to

$$e(D) = \sum_{n=0}^{L-1} r^2(n) - \frac{\left(\sum_{n=0}^{L-1} r(n)r(n-D)\right)^2}{\sum_{n=0}^{L-1} r^2(n-D)} \tag{2.7}$$

Minimising the error, *e(D)*, is equivalent maximising the second term of Equation (2.7). The search is performed over the range of pitches, 20-147 samples (128 delays to facilitate 7-bit pitch encoding) and the optimum delay, D, is found. For periodic signals, D is the pitch period (or multiple of the pitch period). For non-periodic signals, D has an erratic behaviour. The gain, λ, signifies the level of periodicity, with λ approaching 1 for very periodic signals.

A single-tap pitch synthesis filter can be expressed as:

$$\frac{1}{P(z)} = \frac{1}{1-\lambda z^{-D}}. \tag{2.8}$$

Performance may be improved with multi-tap pitch predictors, however, this is at the expense of increased complexity and coding rates. As an alternative, pitch filters which refine the pitch estimate to a fraction of a sample have been proposed [Kroo90] [Marq90]. These require only a minor increase in the bit rate.

2.8. Linear Prediction-based Analysis-by-Synthesis (LPAS) Coders

Linear Prediction-based Analysis-by-Synthesis (LPAS) coders reproduce speech by passing an excitation signal though a time-varying synthesis filter, the coefficients of

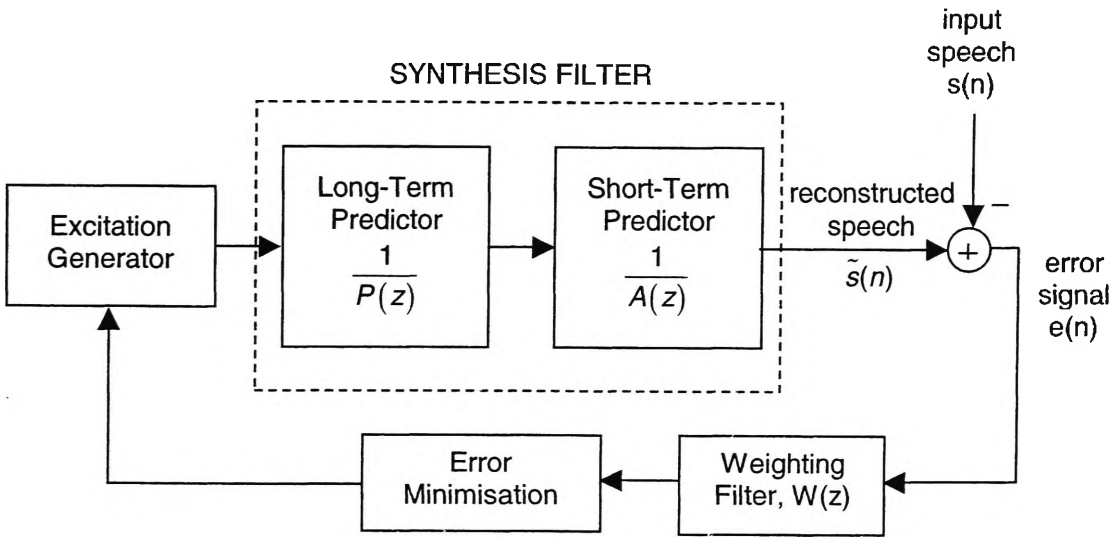


Figure 2.6. Linear prediction-based analysis-by-synthesis technique with perceptual weighting of the error

which are determined by LP analysis. The key element of this technique resides in the method in which the excitation signal is selected. Rather than coding the excitation directly, the appropriate excitation is chosen by using a closed-loop optimisation process. This allows direct analysis of the reconstructed speech. Candidate excitation segments are filtered with a replica of the synthesis filter during analysis, and the one that minimises a perceptually weighted distortion measure, usually the Mean Squared Error (MSE), is selected. The analysis-by-synthesis technique can successfully lower the bit rate, as it provides a means to check the coded parameters, as well as account for errors from previous frames. This results in the best excitation from the available selection to be chosen. A block diagram of LPAS is shown in Figure 2.6.

Perceptual Weighting

Perceptual weighting of the error signal incorporates auditory masking effects into the minimisation calculation and attempts to optimise performance to the human ear [Atal79]. The weighting emphasises the error in spectral valleys, and de-emphasises the error near formant peaks. This shapes the quantisation noise such that it is less audible due to masking by the high energy parts of the signal.

The weighting filter is given by:

$$W(z) = \frac{A(z)}{A\left(\frac{z}{\gamma}\right)} = \frac{1 - \sum_{k=1}^{P} a_k z^{-k}}{1 - \sum_{k=1}^{P} a_k \gamma^k z^{-k}} \qquad 0 \le \gamma \le 1 \tag{2.9}$$

where the factor γ controls the amount of de-emphasis.

2.8.1. LPAS Implementations

Most of the early LPAS methods used only a short-term filter, $\frac{1}{A(z)}$, for synthesis, but later included a long-term (pitch) filter, $\frac{1}{P(z)}$, which improved performance significantly, especially for female speakers. A common early implementation of LPAS is Multi-pulse Excitation (MPE), proposed by Atal and Remde [Atal82], in which the excitation is represented as a sequence of non-uniformly-spaced pulses. The coder determines both the location and amplitude of each pulse during

analysis, such that the weighted MSE is minimised. A 9.6kbits/s MPE algorithm was adopted for Skyphone airline applications. Also, the lower complexity Regular Pulse Excitation (RPE) was developed by Kroon *et al.* [Kroo86], in which regularly-spaced pulsed patterns were used for the excitation. This simpler model required only the initial pulse position and the amplitude of each pulse to be transmitted. A modified version of RPE which incorporated long-term prediction was adopted for the full rate GSM Pan-European digital mobile standard. Both MPE and RPE produce good quality at medium bit rates, but the individual quantisation of the pulse parameters is inefficient for lower bit rates. This led to the development of the most notable of the LPAS coders, the Code-Excited Linear Prediction (CELP) coder, attributed to Atal and Schroeder [Atal84][Schr85].

2.8.2. Code-Excited Linear Prediction (CELP)

In the CELP coder, coding efficiency is significantly increased by employing vector quantisation (VQ) [Gers91] to code the excitation sequence, rather than scalar quantisation. This allows the excitation of a subframe to be represented with fewer bits. For example, in [Schr85], a codebook size of 1024 vectors of length 40 samples, requires only 10 bits to represent the shape of the excitation, plus an additional 5 bits to represent the gain parameter, using a 32-level non-uniform scalar quantiser. In CELP, excitation vector candidates are stored in codebooks which are available at both the encoder and decoder. For each input segment, the encoder performs an exhaustive search through the codebook to find the vector which will best represent the current excitation sequence, i.e. produce the smallest error between the original

and reconstructed sequences. The codevector index is then transmitted to the decoder. In most current CELP coders, the excitation vector typically consists of a contribution from an adaptive codebook [Klei88] and a stochastic codebook. The adaptive codebook provides the correct periodicity of the speech, which is important for good perceptual speech quality. It is made up of shifted and scaled, overlapping previous excitation segments, and is an alternative structure for the long-term (pitch) filter. The difference between the input excitation and the chosen adaptive codebook contribution is then used during the search of the fixed (stochastic) codebook contribution, which is populated with Gaussian sequences. The original CELP algorithm required high computational complexity for the codebook search, motivating subsequent improvements [Davi86][Kleij90][Gers90]. A block diagram of the CELP encoder with two codebooks is shown in Figure 2.7, where the position of the weighting filter has been shifted to save computational effort. A post-filter, applied to the reconstructed speech, can enhance the speech quality of CELP coders by emphasising the formant structure and making the quantisation noise in the spectral valleys less audible [Chen87].

CELP algorithms are especially successful in the range of bit rates from 4-16kbits/s and, as a result, many speech coders based on CELP have been selected as ITU speech coding recommendations. Some examples are G.728, a 16 kbit/s low-delay CELP (LD-CELP) coder [Chen92], G.729, an 8 kbit/s conjugate-structure-algebraic CELP coder (CS-ACELP) [Sala98] and FS-1016, a 4.8kbits/s CELP coder for secure telephony [Camp89].

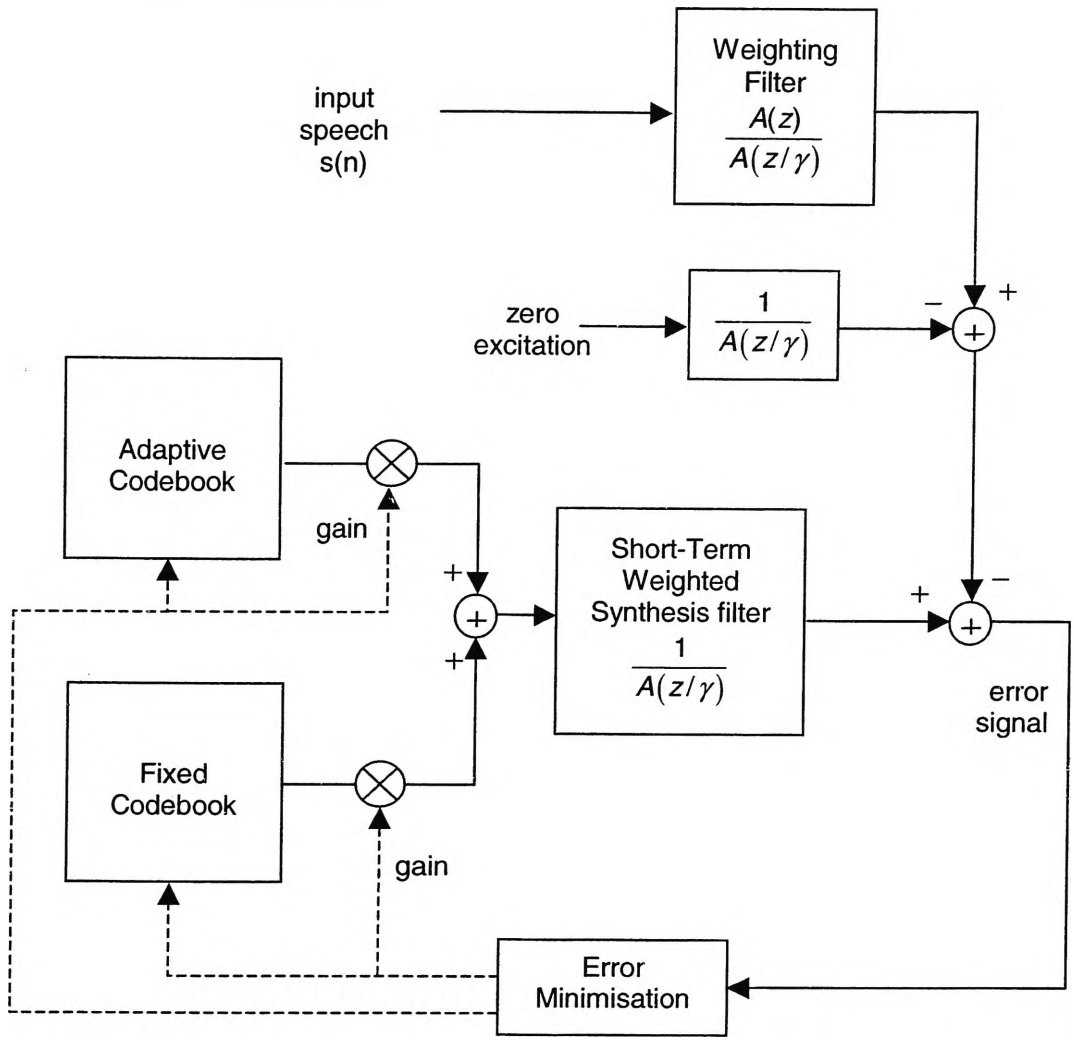


Figure 2.7. CELP encoder with an adaptive codebook and a fixed codebook

Considerable work today is focussed on improving and extending the CELP algorithm. The main areas of research are in adapting CELP for wideband speech coding [Schn98][Kois98][Comb99], variable-rate implementations for the ETSI adaptive multi-rate (AMR) standard [Ekud99][Paks99][Hein99], improving the performance in background noise [Hage99][Beri99], and faster search procedures [Ha99][Rami99].

2.8.3. Generalised Analysis-by-Synthesis

A disadvantage of the error minimisation process of analysis-by-synthesis, is that small phase delays, which are perceptually insignificant, result in large matching errors. Hence, generalised analysis-by-synthesis was introduced by Kleijn *et al.* [Klei92] to account for the insensitivity of the human ear to certain distortions. In generalised analysis-by-synthesis, a modifier is incorporated into the analysis-by-synthesis structure. The modification to the input signal must be perceptually insignificant, but it creates changes to the signal which result in a significant reduction in the required bit rate. This principle was the basis for the Relaxed CELP (RCELP) implementation proposed by Kleijn *et al.* [Klei93], which increase the efficiency of coding the pitch parameter. In these coders, the pitch is estimated once per frame, then linearly interpolated. The LP residual is then modified by time-warping (or time-shifting) to be consistent with the pitch contour, and CELP techniques are performed on the modified residual signal.

2.9. Frequency Domain Coders

Instead of coding the full bandwidth of the signal at once, greater coding efficiency can be achieved by decomposing the signal into a number of frequency subbands, either by a filter bank (subband coding) or by a transform (transform coding). The subbands are then separately encoded. This allows explicit control of each frequency region, and the ability to exploit perceptual knowledge and signal redundancies in the frequency domain.

2.9.1. Subband Coders

In subband coders, e.g. [Cox88], the signal is divided into typically 4-16 frequency subbands, often by using uniform M-channel filter banks (Figure 2.8) or tree-structured (wavelet) filter banks (Figure 2.9). Uniform filter banks split the spectrum into M frequency subbands of equal width. The filters $f_0(n), f_1(n),\ldots, f_M(n)$ are often modulated versions of a prototype filter [Vaid93]. Tree-structured filter banks split the spectrum into a low-frequency band and a high-frequency band. Both bands are down-sampled by 2, and the output of the lower frequency band is then further split. This results in a non-uniform, dyadic frequency separation which is similar to the critical bands associated with the human ear. It is also observed that the formants of speech are much narrower at lower frequencies, further justifying the spectrum division to have smaller bandwidths at low frequencies and wider bandwidths at high frequencies. Of most interest are filter banks which satisfy the perfect reconstruction property, i.e. aliasing and phase distortions incurred due to the overlapping of subbands of the analysis filters are cancelled out by the synthesis filters. The most commonly used of these filter banks are two-channel biorthogonal quadrature mirror filter (QMF) banks [Stran96], with either finite impulse response (FIR) or infinite impulse response (IIR) filters.

The reduced bandwidth of the subband signals means they can be downsampled prior to encoding. Subband coders have the advantage of containing quantisation noise within the band, preventing it from masking or interfering with an adjacent frequency range. In addition, if PCM is used to quantise the subbands, the

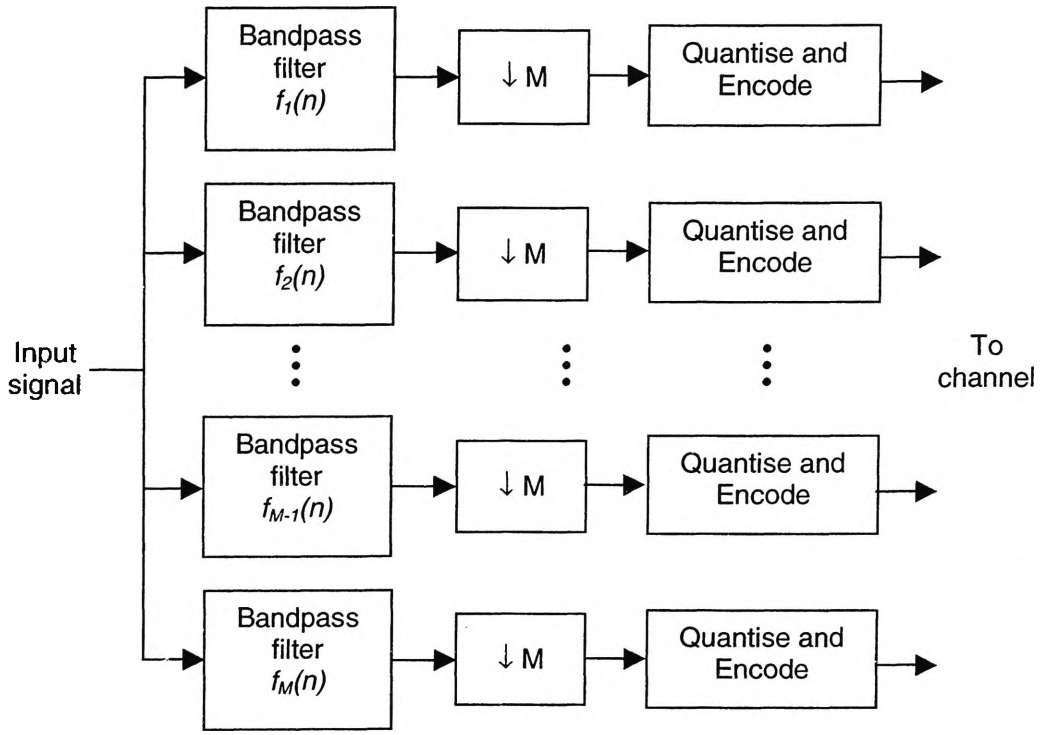


Figure 2.8. Division of the input signal into frequency subbands using a uniform M-band filter bank

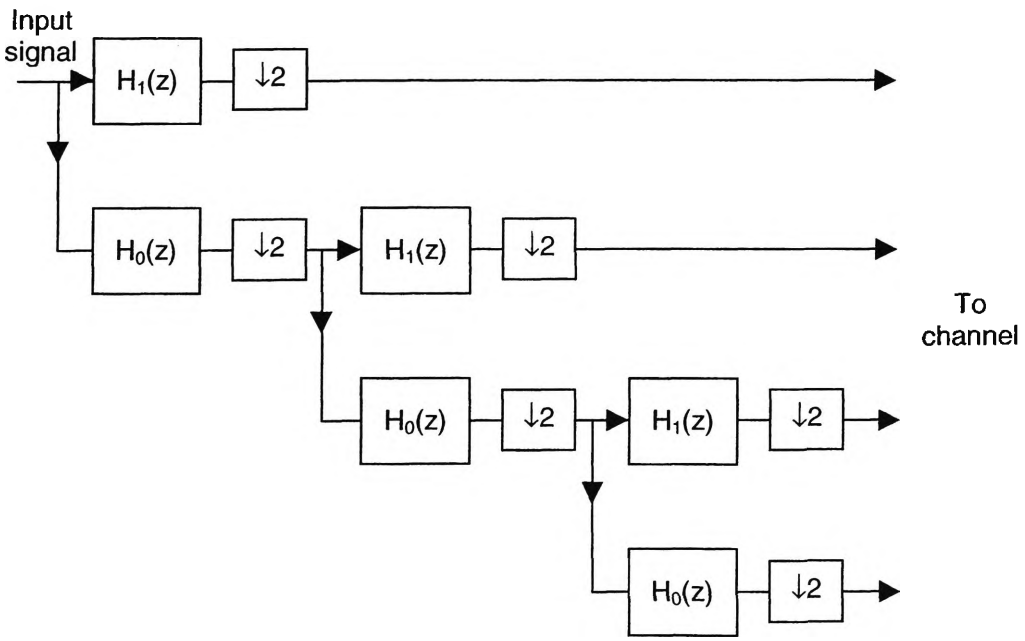


Figure 2.9. Division of the input signal into frequency subbands using a tree-structured filter bank

quantisers may exploit the non-flat nature of the signal spectrum and adaptively adjust the dynamic range and step-size of the quantisers to match the energy of the particular band. They may also incorporate perceptual criteria, preferentially allocating bits to different frequency bands. For the coding of speech, the lower frequencies are often allocated more bits to preserve critical pitch and formant information. More sophisticated subband coders may further improve coding efficiency by taking advantage of closely related wavelet theory, e.g. optimisation of the filter bank design to better match the human auditory system [Prin95].

2.9.2. Transform Coders

Transform coders, e.g. [Trib79][Zeli79], possess similar advantages to the subband coder, and perform short-time spectral analysis of speech by using block transformations of windowed sections of speech. They provide a "narrow-band" analysis dividing the input signal into typically 64-512 frequency components. Efficiency is gained through assigning more bits to the important transform coefficients than to the less important coefficients. Block transforms do, however, suffer from edge effects (discontinuities at the block boundaries), which may be reduced by using overlapping windows. The disadvantage is that the overlapped frequency bands generate aliasing, causing a spreading of energy and worsening of quantisation errors which were not cancelled during synthesis. Transforms often used are the Discrete Cosine Transform (DCT), which has near-optimal performance to the Karhunen-Loève Transform, the Discrete Fourier Transform (DFT), and the Walsh-Hadamard Transform .

Subband and transform coders produce high quality speech at rates around 16 kbit/s, but the output quality is poor when the bit rate drops to 8kbits/s. This can be attributed to the absence of a pitch predictor, and hence inadequate reproduction of the speech periodicity.

2.9.3. The Pitch-Synchronous Wavelet Transform

Wavelet and subband coding techniques have demonstrated advantages in signal compression. A further extension of this idea is the Pitch-Synchronous Wavelet Transform (PSWT), introduced by Evangelista [Evan93], which has the additional favourable attribute of exploiting the pitch periodicity of speech signals. The PSWT operates across a pitch-synchronous (PS) representation of speech, which consists of sequences of pitch-length segments (pitch periods), during voiced speech, and reduces to the signal itself during unvoiced speech. The PSWT performs a series of wavelet transforms over sections of the PS representation, i.e. over samples spaced one period apart. Hence, for periodic segments, whereas, the ordinary WT represents the signal in terms of a trend (local average) plus fluctuations at several scales, the PSWT represents the signal in terms of a regular periodic component plus period-to-period fluctuations at several scales. This is depicted in Figure 2.10 For aperiodic segments, the PSWT reduces to the ordinary Discrete Wavelet Transform (DWT). However, application of the PSWT in real-time speech coding applications is not discussed in [Evan93], and an adaptation of the method to perform the decomposition in real-time is described in Chapter 3. In addition, a novel technique

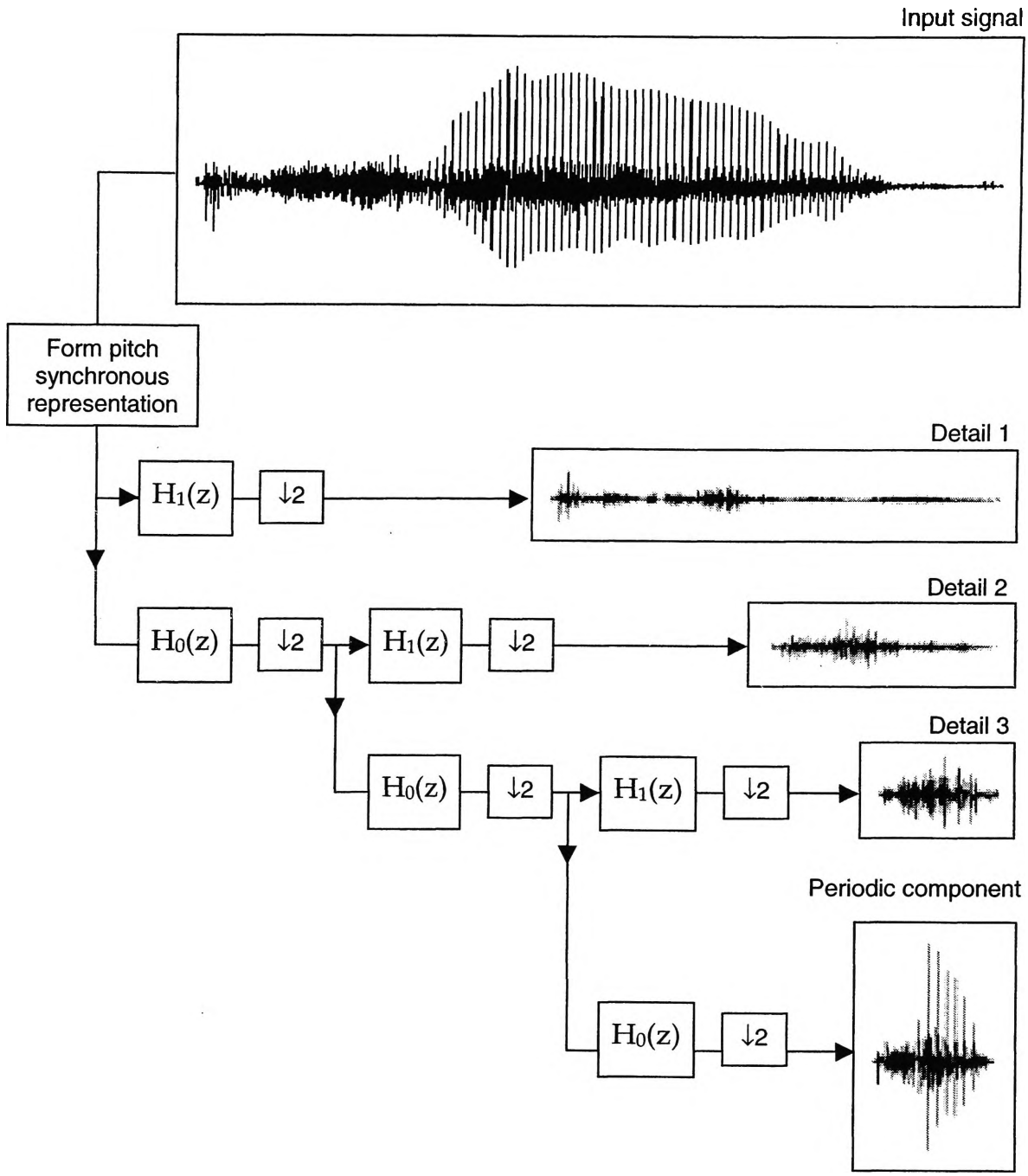


Figure 2.10. Decomposition of the LP residual by the PSWT into a low-resolution periodic component plus period-to-period fluctuations (detail) at several scales. The filters H_0 and H_1 are half-band lowpass and highpass respectively.

to perform the required pitch-synchronous extraction of pitch periods required for the PSWT is presented in Chapter 4.

An advantage of the PSWT is the ability to perform good separation of vowel sounds from unvoiced fricatives and nasal consonants, especially if the pitch is not changing rapidly. This facilitates the ability to code voiced (periodic) sounds with highest priority, since they contain the most perceptually relevant information, followed by the wavelet partials (fluctuations) in decreasing scale order. The wavelet partials add dynamics and naturalness to the sound.

2.10. Sinusoidal Coders

2.10.1. Sinusoidal Transform Coders

Sinusoidal models are successful in low-rate speech coding as they exploit speech periodicity. In the sinusoidal speech model proposed by McAulay and Quatieri [McAu95], speech may be represented as a linear combination of L sinusoids with time-varying amplitudes, phases and frequencies. This sinusoidal representation is described by

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) \tag{2.10}$$

where A_l, ω_l, ϕ_l, represent the amplitude, frequency and phase of each sine-wave component.

Since voiced speech is highly periodic, coding gains can be achieved by constraining the set of sinusoids, e.g. limiting the frequencies to be integer multiples of the pitch. Hence, equation (2.10) becomes

$$s(n) = \sum_{l=1}^{L} A_l \cos(nl\omega_0 + \phi_l), \qquad n = 1,2,3,... \tag{2.11}$$

where ω_0 is the pitch frequency, and A_l, ϕ_l, are determined by the Short-time Fourier Transform (STFT). For unvoiced speech, the frequency separation between the sinusoids is small, such that the power spectrum changes slowly over consecutive frequencies.

2.10.2. Multi-band Excitation (MBE) Coders

A variation of sinusoidal coding is Multi-band Excitation (MBE) coding proposed by Griffin and Lim [Grif88]. In this method, the pitch of the frame is first estimated and the spectrum is divided into subbands by the Fast Fourier Transform (FFT). A synthetic spectrum is formed using the harmonic magnitudes and the pitch frequency. Each subband is declared as either voiced or unvoiced, depending on the similarity between the original and synthetic spectrum, allowing a mixture of both harmonic and random contributions to model the spectrum. An Improved MBE (IMBE) coder operating at 6.4 kbit/s was developed by Hardwick and Lim [Hard91], and incorporated more efficient techniques to code the parameters of the MBE analysis-by-synthesis model. This coder was adopted as a standard for satellite voice communications by Inmarsat, and DVSI (Digital Voice Systems, Inc, www.dvsinc.com) have continued to develop a series of coders based on IMBE, one of which was adopted for the ill-fated Iridium project.

2.11. Waveform Interpolation Coders

2.11.1. Prototype Waveform Interpolation (PWI)

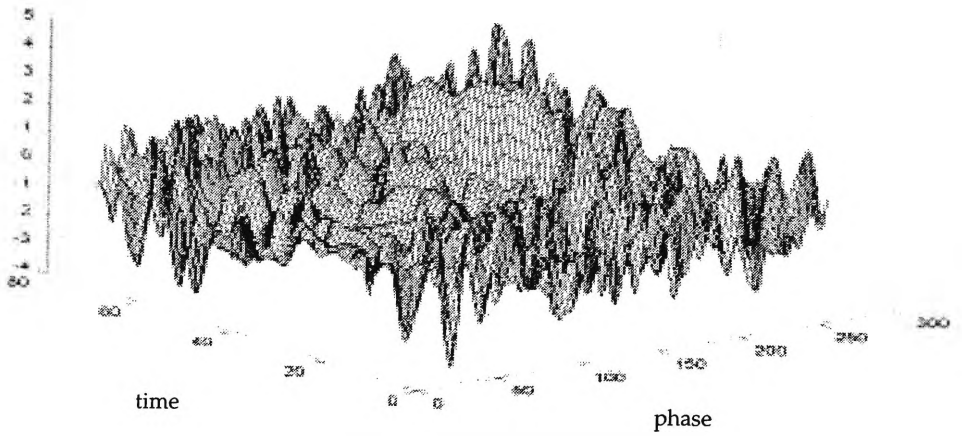
The Waveform Interpolation (WI) principle, proposed by Kleijn [Klei93b], is based on the observation that voiced speech can be interpreted as a sequence of pitch-cycle waveforms that evolve slowly over time. This suggests that voiced speech has a relatively low information content and may be accurately reproduced by prediction and interpolation of sparsely located representative pitch cycles, called prototype waveforms. Prototype Waveform Interpolation (PWI) was motivated by the concept that the perceptual quality of voiced speech in CELP can be improved by increasing its periodicity, even though the SNR reduces [Shoh91]. Hence, PWI aims to preserve the level of periodicity in voiced speech, and maintain high perceptual quality, even at low rates. During voiced sections, a single pitch cycle, the prototype waveform, is transmitted at regular intervals (every 20-30ms). The instantaneous excitation can then be obtained by linearly interpolating between the transmitted waveforms, producing a reconstructed waveform with a similar shape to the original speech, but with loss of time-synchrony. The method was initially combined with a CELP coder for the coding of unvoiced speech, obtaining good quality at around 3kbit/s [Klei93b][Burn93].

2.11.2. Waveform Interpolation (WI)

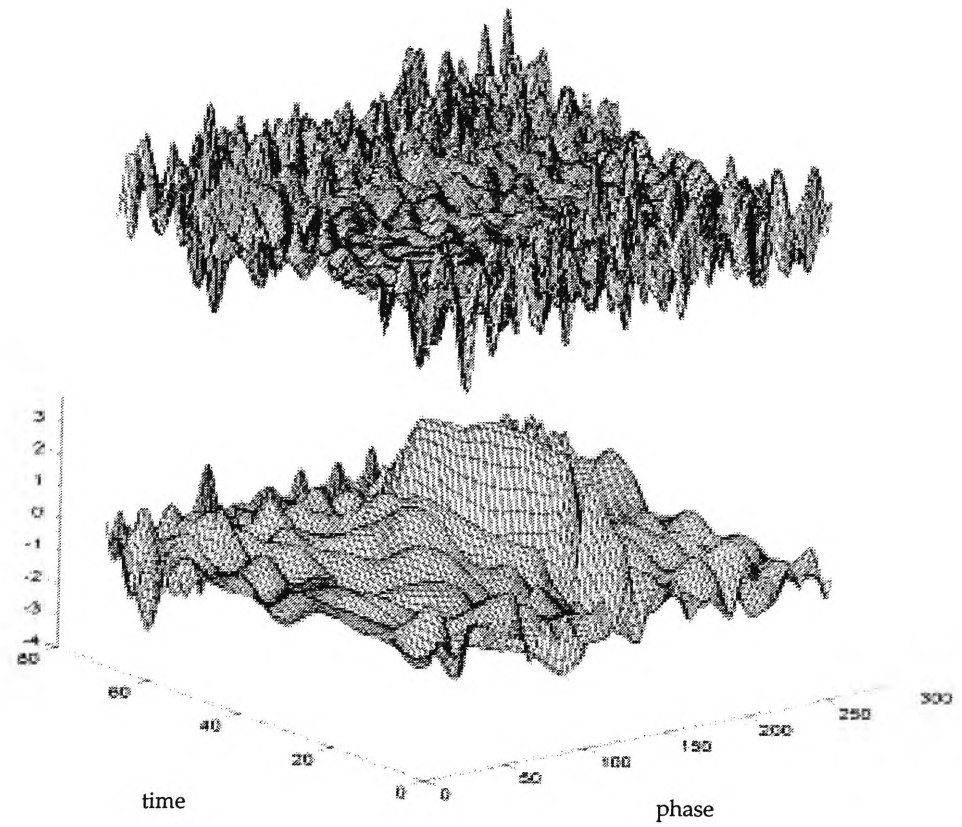
The PWI paradigm was later extended to code both voiced and unvoiced speech by increasing the extraction rate of the prototype waveforms (now referred to as characteristic waveforms, CW) [Klei94]. The higher extraction rate facilitates the

coding of unvoiced signals which do not display periodicity. While this also implies a higher bit rate, efficient quantisation of the signal is achieved by using a transformation and decomposition. The one-dimensional speech signal is transformed into a two-dimensional surface with the CW shape along one axis, and the evolution of this shape over time on the other axis. The CW surface evolves slowly during quasi-periodic, voiced speech and rapidly during noise-like, unvoiced speech. Thus, the signal can be decomposed into a periodic and non-periodic component, the slowly-evolving waveform (SEW) and rapidly evolving waveform (REW) respectively, by simple lowpass and highpass linear filtering the CW surface evolution. The gain-normalised CW surface and the decomposed SEW and REW surfaces are shown in Figure 2.11. These components, which are perceived very differently by the human auditory system, can then be separately and efficiently quantised in a perceptually accurate manner. Simplified block diagrams of the WI analysis and synthesis are shown in Figure 2.12 and Figure 2.13 respectively. The high level of computational complexity of WI can be reduced using techniques described in [Klei96].

A variation of the WI technique is the Pitch Pulse Evolution (PPE) model proposed by Stachurski and Kabal [Stac94][Stac98]. In this approach, pitch pulses are extracted pitch-synchronously, as in [Klei96], rather than at a fixed rate. This makes the WI decomposition by simple fixed-coefficient, linear filtering less effective due to the need for a variable-tap filter. A complex decomposition procedure is proposed which estimates the underlying pitch pulse shape (analogous to the SEW) from the excitation signal, by minimising a weighted error criterion, using Singular



(a) Normalised Characteristic Waveform Surface



(b) Rapidly Evolving (upper) and Slowly Evolving Surfaces

Figure 2.11. Decomposition of the evolving CW surface into a slowly evolving component and a rapidly evolving component

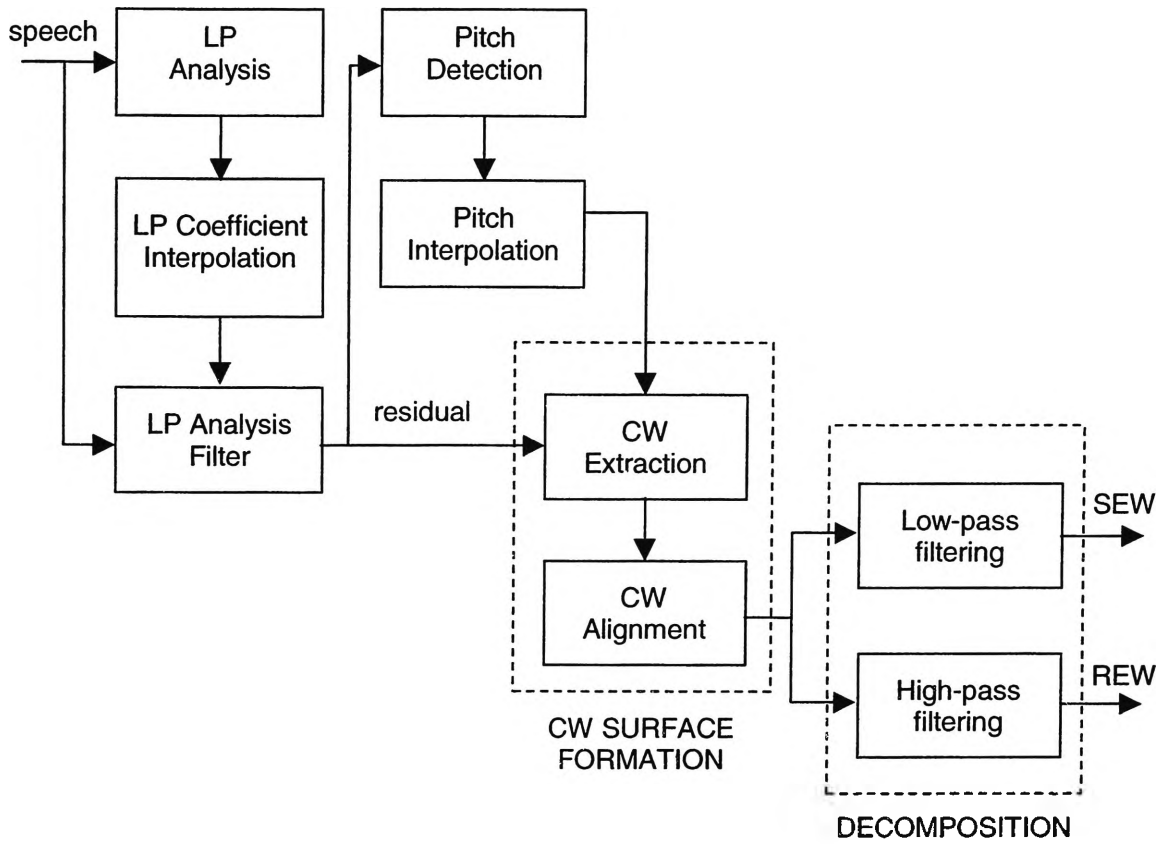


Figure 2.12. Simplified block diagram WI analysis.

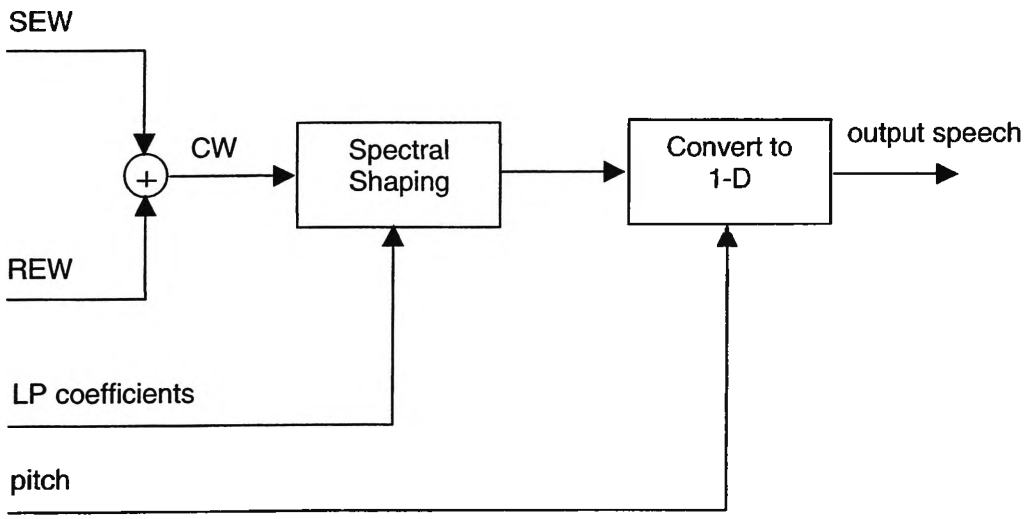


Figure 2.13. Simplified block diagram WI synthesis

Value Decomposition theory. The technique also attempts to maintain approximate time-synchrony between the input and output signals, placing importance on the pitch pulse locations. While initial comparison tests without the pulse shape quantised produced good results, the coder failed to accurately represent the very perceptually significant underlying pulse shape.

Techniques to improve the quantisation of the characteristic waveforms of WI include the use of closed-loop linked codebook structures [Burn95], analysis-by-synthesis techniques [Burn97], multiband voicing analysis [Yagh97], and perception-based techniques which exploit the frequency resolution of the human ear [Thys97]. Recent work in WI has focussed on adapting WI to increase the scalability and flexibility of the coder and obtain good quality at higher rates. An embedded WI coder, operating in the range of 2.0 - 4.8 kbit/s was proposed by [Kang99b]. In this approach, the transmission frequency and bit allocations for the SEW and REW components adapts, depending on the bandwidth availability per user. Similarly, scalability can be obtained within perfect reconstruction frameworks.

2.11.3. Perfect Reconstruction Waveform Interpolation

Some of the methods used in conventional WI are not amenable to perfect reconstruction, such as the circular pitch-cycle alignment of overlapped, pitch-length extracted waveforms. Hence, Yang and Kleijn [Yang98] adapted the WI principle to achieve good performance at low bit rates, as well as convergence to perfect quality with increasing bit rate. The modified WI coder uses a time-warping

of the residual signal to facilitate pitch-synchronous, fixed-length analysis. However, accurate time-warping of the signal is crucial for good performance of the paradigm, and a reliable technique to form a pitch track, which ensures the extracted warped pitch periods are phase-aligned, was not described. The approach also applies fixed-length transforms, such as the Gabor Transform (GT) and the Modulated Lapped Transform (MLT), implemented as perfect reconstruction filter banks, to produce parameters which are more easily quantised. The transforms, however, do introduce significant delays, and in the case of the GT, redundancy.

More recently, an approach using a similar principle to that of RCELP was presented to also incorporate perfect reconstruction properties into the WI paradigm [Erik99]. This work was performed concurrently with the early work of Chapter 4 of this thesis [Chon99b]. It thus provides an interesting, contrasting approach to the perfect reconstruction problem. Instead of time-warping, small and large time-shift modifications are made to the residual, to establish a modified signal with the desired properties for transformation and coding. The disadvantage of this technique is that some samples (and almost whole pitch periods) of the original signal may be omitted or repeated in the modified signal, which could result in noticeable distortions when the errors accumulate.

An ongoing area of interest in WI is the issue of phase representation. A number of techniques to improve upon traditional phase models have been proposed recently. These include an analysis-by-synthesis vector quantisation scheme for the SEW phase [Gott99], which uses multiple codebooks to accommodate varying pitch

period values, and also a method for adjusting the REW spectrum based on the variation of the pitch and the audibility of noise at particular frequencies [Kang99]. These methods emphasise the importance of correctly representing phase information to obtain natural sounding speech, and, in the case of [Gott99], aim towards improving waveform matching.

2.11.4. Waveform-Matched Waveform Interpolation

The Waveform Matched Waveform Interpolation (WMWI) approach described in this thesis, enables improved speech analysis over existing WI coders through producing an accurate representation of speech evolution, as well as facilitating waveform coding. This approach uses a reliable technique for estimating the fundamental frequency of the speech and refines this estimate to form an optimised pitch contour. The pitch track is designed to warp the speech residual to have a constant pitch, and such that its pulse peaks occur at the central sample of each pitch period. This allows pitch-synchronous extraction of fixed-length pitch cycles, and effective evolutionary waveform decomposition. The WMWI analysis techniques are invertible, without causing errors due to cyclic rotation, or the repetition or omission of samples due to selective extraction. Hence, near-perfect reconstruction of the input signal can be achieved. Effective methods for pitch quantisation and reconstruction are also applied to preserve time-synchrony between the input and output signals.

2.12. Summary

In this chapter, a number of current speech coding technologies have been discussed. The processes of speech production and perception were briefly described; an understanding of which is essential in order to develop efficient speech coding algorithms. The discussion began with the PCM coder, which produces the simplest digital representation of a waveform, then moved on to introduce methods which exploit speech redundancies to achieve bit rate reductions. The most widely used techniques include linear prediction, to remove the short-term correlations (the spectral envelope) and pitch prediction, to remove the long-term correlations (the spectral fine structure).

The review of existing speech coders has shown a positive trend towards decomposing speech into components which can be effectively quantised. For example, in subband or transform coders, the spectrum is divided into multiple frequency subbands allowing separate and adaptive bit allocations, and in WI coders, the evolution of pitch-cycle waveforms is separated into a slowly-evolving and rapidly evolving component, enabling the different perceptual qualities of voiced and unvoiced speech to be exploited.

We can also notice a drive towards developing coders which are scalable in bit rate, producing good quality speech at both low and medium rates. These coders aim to bridge the gap between strictly waveform and strictly parametric coders by incorporating models of speech production and perception with the property that the output performance converges to perfect quality with increasing bit rate. The

most widely recognised and studied of these coders is the CELP coder, which combines the highly effective linear predictive analysis-by-synthesis technique with the efficient method of vector quantisation. Analysis-by-synthesis allows an improved selection of the excitation sequence by incorporating the synthesis procedures in the analysis loop, and thus enabling direct feedback on the consequence of quantisation errors. The quality of CELP coders, however, degrades significantly when the bit rate drops below 4kbits/s, as the codebook size becomes too small to effectively preserve significant pitch and formant information.

Experiments testing the perceptual sensitivity of the human auditory system have revealed the importance of maintaining the correct degree of periodicity in speech. This has stimulated approaches which exploit the periodic structure of speech. Sinusoidal coders, at low rates, concentrate on quantising the amplitudes, phases and frequencies of only the pitch harmonics. Alternatively, the WI paradigm, which has gained a great deal of attention in recent years, is based on the evolution of pitch-length, characteristic waveforms. Such a representation enables reconstruction by interpolation. Pitch-synchronous approaches have also been proposed, including the PPE model and several modified WI techniques. These techniques have a variable-rate analysis, although in some cases, the analysis segment length is made constant by time-warping, allowing fixed-length analysis. In addition, the analysis and synthesis procedures of modified WI coders have been designed to achieve the waveform coding property.

We have thus identified the favourable components found in speech coders to produce high quality speech at low bit rates. An ideal coder should accurately represent the spectral properties of speech, decompose the signal into components with different quantisation requirements, optimise performance to the human ear, and preferably, provide for speech waveform matching. These properties are incorporated in the Waveform-Matched Waveform Interpolation (WMWI) coder proposed in Section 4. The main features of WMWI include the reliable transformation of the speech residual into a surface of critically sampled, phase aligned warped pitch periods, invertible analysis procedures, and time-synchronous reconstruction of the quantised speech.

While the focus of the chapter was on methods to reduce the bit rate while maintaining high output quality, there are many other important properties of a speech coder which have not been discussed here. These include coder complexity and cost, memory requirements, delay, and coder robustness. The emphasis on each of these properties will depend on the various target applications and their quality objectives.

Chapter 3

Encoding Speech Evolution Using Wavelet Decomposition

An idea isn't responsible for the people who believe in it.

-- Don Marquis

3.1. Introduction

In recent WI coders, simple linear filtering along the evolution of the speech residual decomposes the signal into two components: a quasi-periodic, voiced component and a non-periodic, noise-like component [Klei94]. The decomposition allows efficient quantisation of the signal by coding the decomposed waveforms with an accuracy based on their perceptual importance. This exploits knowledge of how the human ear perceives different types of sounds. It is then anticipated that a decomposition which produces multiple evolution frequency subbands, will enable further improvement in the quantisation efficiency and synthesised speech quality. In addition, for the coding of speech in noisy environments, a frequency band analysis may be advantageous to isolate and suppress unwanted noise components,

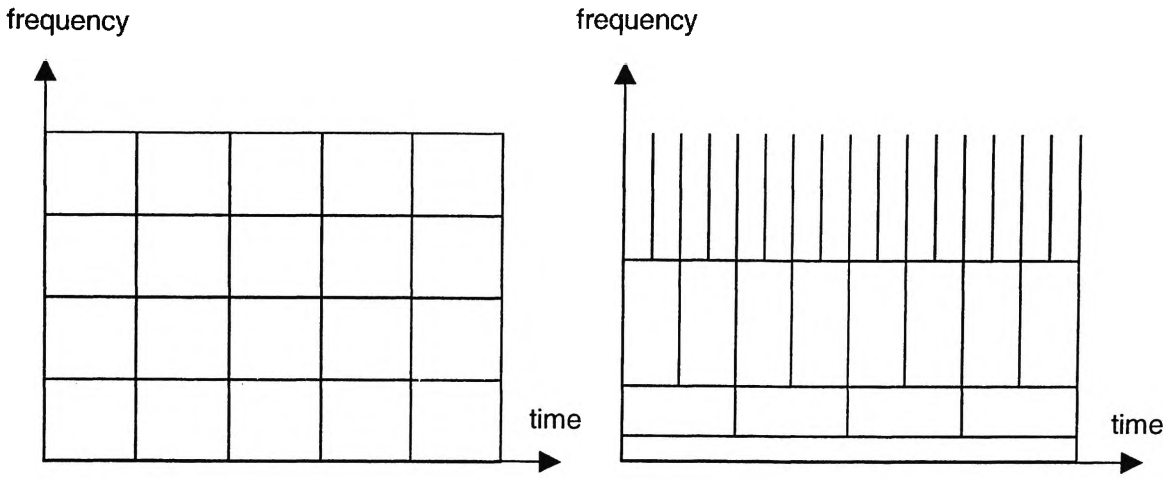
the effects of which were studied in [Chon97] for the standard WI decomposition. This motivated the desire for an alternative decomposition method for WI coding.

A solution to multi-evolution frequency subbands lies with a decomposition based on the Pitch Synchronous Wavelet Transform (PSWT) [Evan93], which may be implemented as a quadrature mirror filter (QMF) bank. When incorporated in the WI framework, the PSWT is able to decompose the characteristic waveform surface into a series of component surfaces, each representing a particular evolution frequency subband. This provides an alternative, more detailed, description of signal evolution – a time-scale representation. The PSWT also offers significant advantages due to the flexibility in the design of its perfect reconstruction filter banks, a strong basis for which has been formed in wavelet and filter bank theory [Vett92][Stran96].

In this chapter, the PSWT is introduced and is integrated into the WI coder. Several finite impulse response (FIR) and infinite impulse response (IIR) wavelet filter designs which achieve perfect reconstruction are presented and a preferred technique for the quantisation of the decomposed surfaces is identified.

3.2. The Discrete Wavelet Transform (DWT)

The Wavelet Transform (WT) provides an alternative to the traditional Short Time Fourier Transform (STFT) for the analysis of non-stationary signals. In contrast to the STFT, which uses a fixed analysis window, the WT uses short analysis windows at high frequencies and long analysis windows at low frequencies, providing a non-



(a) STFT (b) WT

Figure 3.1. Time-frequency planes for the STFT and WT

uniform coverage of the time-frequency plane (Figure 3.1). This produces a multi-resolution representation of the signal, with dyadic localisation in time and frequency, as opposed to the fixed resolution of the STFT. This is the essence of the WT. Detailed wavelet theory may be found in the following references: [Daub92] [Mall98][Vett92], however, only the relevant sections are discussed here.

The Discrete Wavelet Transform (DWT) decomposes the signal onto a set of basis functions, called wavelets, $\psi_{n,m}(k)$, and can be expressed as:

$$x(k) = \sum_{n=1}^{N} \sum_{m=0}^{M} X_{n,m} \psi_{n,m}(k), \tag{3.1}$$

where

$$X_{n,m} = \sum_{k} x(k)\, \psi_{n,m}(k) \tag{3.2}$$

are the transform coefficients, index n represents scale and m represents time shift.

The wavelet sequences can be obtained by scaled and translated versions of the mother wavelet, $\Psi_{n,0}(\omega)$:

$$\psi_{n,0}(k) = IDFT\left\lfloor \Psi_{n,0}(\omega) \right\rfloor, \tag{3.3}$$

$$\psi_{n,m}(k) = \psi_{n,0}(k - 2^n m). \tag{3.4}$$

The mother wavelet can be thought of as a bandpass filter, obtained from the lowpass and highpass filter transfer functions, H_0 and H_1 respectively.

$$\Phi_{n,0}(\omega) = \prod_{k=0}^{n-1} H_0(e^{j2^k \omega}) \tag{3.5}$$

$$\Psi_{n,0}(\omega) = H_1(e^{j2^{n-1}\omega})\Phi_{n-1,0}(\omega) \tag{3.6}$$

3.2.1. Perfect Reconstruction Filter Banks

The DWT can be implemented as a filter bank with discrete-time filters and downsampling by 2. The basic component of a tree-structured filter bank is the two-band analysis/synthesis system depicted in Figure 3.2.

The splitting of the frequency spectrum by the analysis filters $H_0(z)$ (lowpass characteristic) and $H_1(z)$ (highpass characteristic) is shown in Figure 3.3. These frequency responses overlap, resulting in aliasing in each channel. Therefore, in order to recover the signal exactly, the synthesis filters, $G_0(z)$ and $G_1(z)$ must be adapted to cancel the aliasing effects and distortions of the analysis filters.

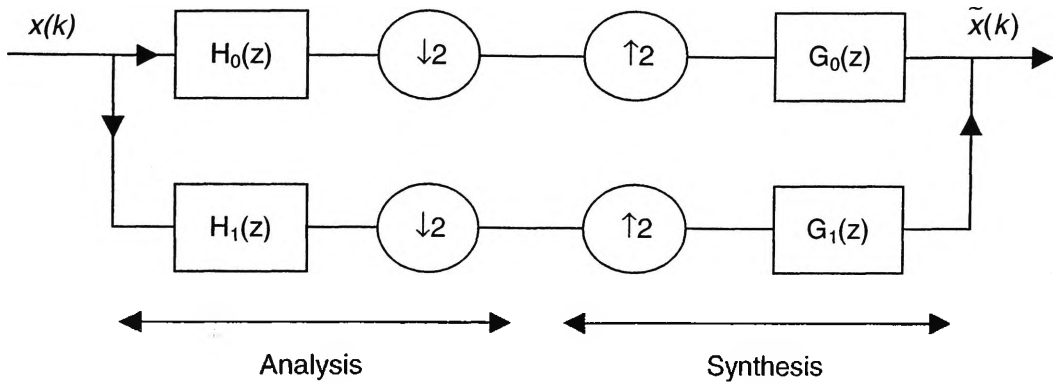


Figure 3.2. Maximally-decimated two-channel filter bank

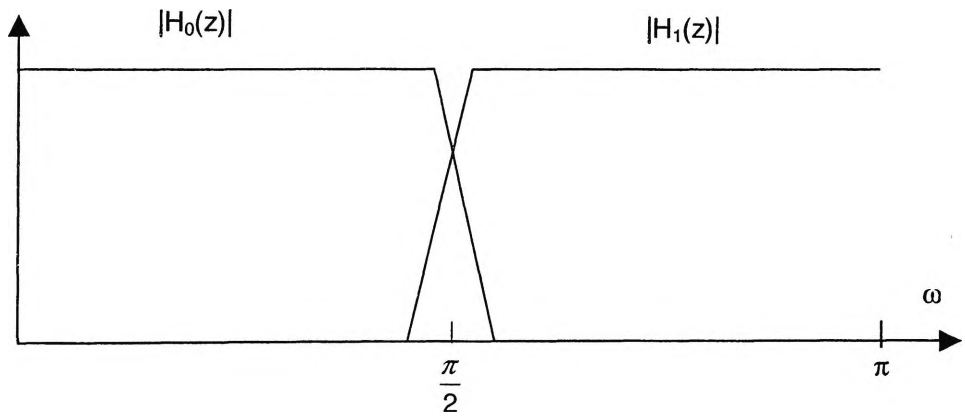


Figure 3.3. Splitting of the spectrum by the filter bank

Alias cancellation is satisfied if the following relation exists (The Perfect Reconstruction Property):

$$G_0(z) = H_1(-z) \tag{3.7}$$

$$G_1(z) = -H_0(z) \tag{3.8}$$

where $H_0(z)$, $G_0(z)$ are scaling sequences (lowpass characteristic) and $H_1(z)$, $G_1(z)$ are wavelet sequences (highpass characteristic).

3.3. Characteristic Waveform Surface Formation

WI is based on a description of waveforms of a single pitch cycle, called characteristic waveforms (CWs). The one dimensional residual signal, $s(t)$, is transformed into a two-dimensional representation, the CW surface, $u(t,\phi)$, which represents the shape of the CW along the ϕ axis, and the evolution of this shape along the t axis. Here, we briefly outline the method by which the CWs are extracted and aligned in WI.

3.3.1. Characteristic Waveform Extraction

CWs are extracted at regular intervals from the linear prediction residual. However, the extraction location is not tightly constrained and may be adjusted to minimise the discontinuities at the segment endpoints due to periodic extension. Early WI coders were only used for the coding of voiced speech, and hence only one CW was extracted per frame. However, for good reconstructed quality of both voiced and unvoiced sounds, a higher extraction rate is required. A typical CW sampling rate is 400 Hz, whereby the CWs are oversampled, though not by a constant rate (due to the natural pitch variations of speech).

3.3.2. Characteristic Waveform Alignment

The extracted CWs are generally not aligned. Phase alignment is however necessary to ensure similar features, e.g. the pitch pulses of adjacent waveforms, $u(t_i,\phi)$ and $u(t_{i+1},\phi)$, occur at the same phase position. This enables the evolution of signal components to be displayed along the t axis, allowing effective quantisation

and smooth interpolation. Alignment is generally carried out in the DFT domain by adjusting the phase of the DFT coefficients for the current CW so that it is aligned with the previous CW. This is achieved by selecting the phase offset, ϕ_i, which maximises the cross-correlation between the waveforms, as shown by the equation:

$$\phi_i = \arg\max_{\phi_i} \sum_{k=0}^{\tau_m - 1} \mathbf{Re}\left[\mathbf{P}_m(k)\,\mathbf{P}^*_{m-1}(k)\,e^{j2\pi k\phi_i}\right] \tag{3.9}$$

where τ_m is the pitch of the m^{th} and current CW, which has DFT coefficients $P_m(k)$. In the time domain, this corresponds to a circular rotation of the pitch cycle.

3.4. The Pitch-Synchronous Wavelet Transform (PSWT)

A pitch-synchronous representation of the input speech signal can be formed by constructing a two-dimensional surface of pitch-cycle, characteristic waveforms extracted from the speech. The surface, shown in Figure 3.4, is formed in a similar fashion to the technique described in Section 3.3, except in this case, CWs are *critically-sampled*, rather than oversampled. A Pitch-Synchronous Wavelet Transform (PSWT) can be performed in the 2-D representation and is equivalent to performing the Discrete Wavelet Transform along the signal evolution. This allows the pitch periodicity of voiced speech to be exploited. The PSWT of the evolutionary waveform, $v_q(k)$, can be expressed by

$$v_q(k) = \sum_{n,m} V_{n,m,q}\,\psi_{n,m}(k), \tag{3.10}$$

where,

$$V_{n,m,q} = \sum_{k} v_q(k)\,\psi_{n,m}(k) \tag{3.11}$$

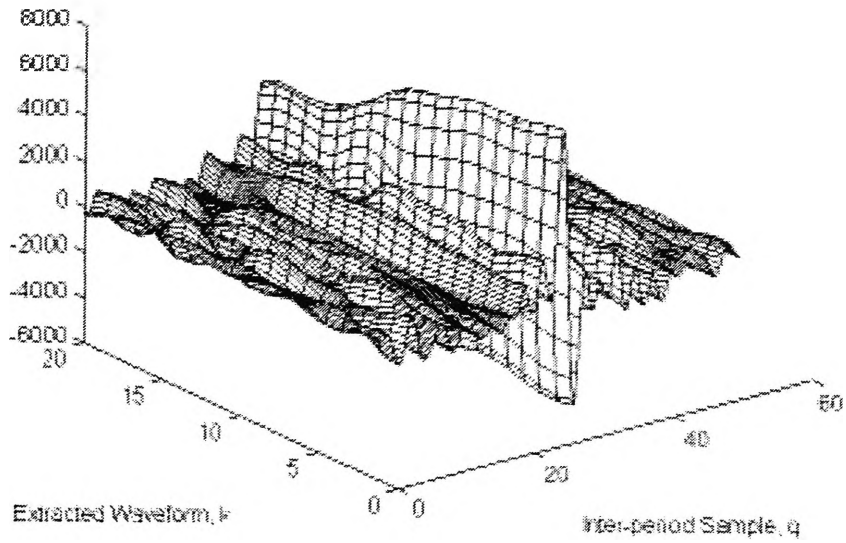


Figure 3.4. Pitch-synchronous representation of the speech residual

are the transform coefficients, and index n=1,2,...,N represents scale and m=0,1,2,...,M represents time shift. Note, an additional index q=0,1,.....*pitch(k)*-1 has been introduced to denote the inter-period sample, along which the transform is being performed.

3.5. Wavelet Decomposition

Strictly, the PSWT requires the pitch cycle waveforms of the CW surface to be critically-sampled, or pitch-synchronous. This requires an extremely accurate pitch detection algorithm to determine pitch lengths and pitch period boundaries, as well as implicitly achieve waveform alignment during the extraction of CWs to enable effective decomposition. A reliable method to perform this in real-time has not been previously established (though a novel solution for reliably extracting pitch periods pitch-synchronously is presented in Section 4.4 of this thesis). Hence, the PSWT is

adapted to maintain a fixed sampling rate, as in WI coding. The rate chosen is 320Hz, corresponding to 2^3=8 CWs per 25ms frame, to provide neat decimation characteristics. The decomposition will now simply be referred to as the wavelet decomposition, as opposed to the PSWT.

The wavelet decomposition is very similar to the SEW/REW decomposition of WI, involving filtering of the signal evolution. This provides for easy integration of the alternative decomposition method into the WI paradigm. The main difference is that the filters are perfect reconstruction wavelets, dilated and shifted to form a non-uniform filter bank (Figure 3.5). Hence, deviations from periodicity are detected in the input speech, isolating the periodic trend and, at the same time, characterising the aperiodic behaviour at several scales. This time-scale analysis provides the potential for improved processing and analysis of the input speech, more sophisticated coding techniques based on perceptual knowledge, as well as feature extraction, for example, to separate the fricative noise from a voiced sound.

The objective of the decomposition is to separate the CW surface into uncorrelated frequency subbands (in the evolution domain). Since each subband signal occupies only a fraction of the original frequency bandwidth it can be downsampled to the Nyquist rate without loss of information. A diagram of the maximally decimated analysis/synthesis system for three decomposition levels is shown in Figure 3.6.

For a N-level decomposition, the final approximation surface,

$$r_N(k,q) = \beta_{N,m,q}(k)\phi_{N,m}(k) \tag{3.12}$$

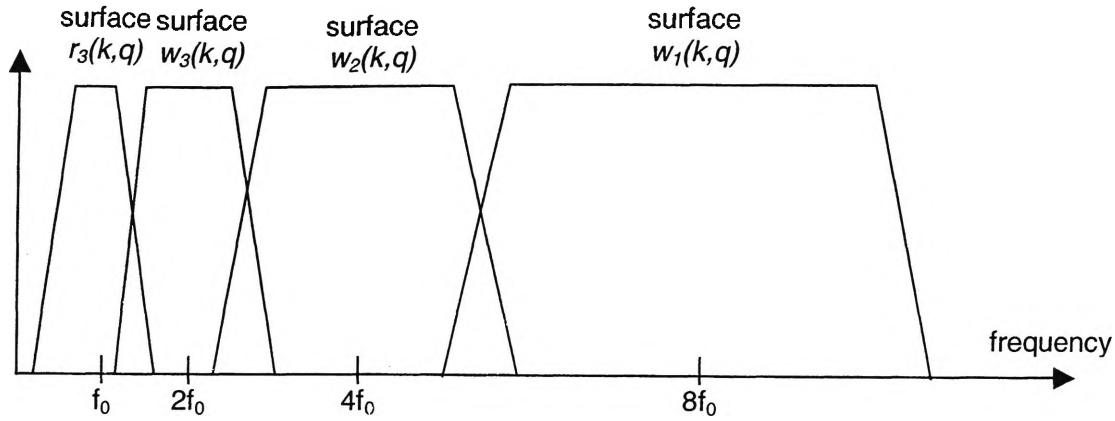


Figure 3.5. Logarithmic coverage of the frequency domain by the WT

with $$\beta_{N,m,q}(k) = \sum_k v_q(k)\phi_{N,m}(k) \tag{3.13}$$

represents the periodic trend, while each detail surface,

$$w_n(k,q) = \sum_m \alpha_{n,m,q}(k)\psi_{n,m}(k) \tag{3.14}$$

with $$\alpha_{n,m,q}(k) = \sum_k v_q(k)\psi_{n,m}(k) \tag{3.15}$$

represents the fluctuations at scale 2^n. The sum of these contributions results in the CW surface:

$$v(k,q) = \sum_{n=1}^{N} w_n(k,q) + r_N(k,q) \tag{3.16}$$

Thus, N detail signals, as well as the approximation signal of the final stage are transmitted (though at different transmission frequencies).

In order to synchronise the signals for reconstruction, extra delays are added in some paths, as indicated in Figure 3.6. The total delay incurred for the decomposition and reconstruction is $z^{(2^N-1)L}$, where L is the combined group

delay of the analysis/synthesis pair. Higher resolution can be achieved by increasing the number of stages, N, in the filter bank implementation, however, this is at the cost of increased system delay.

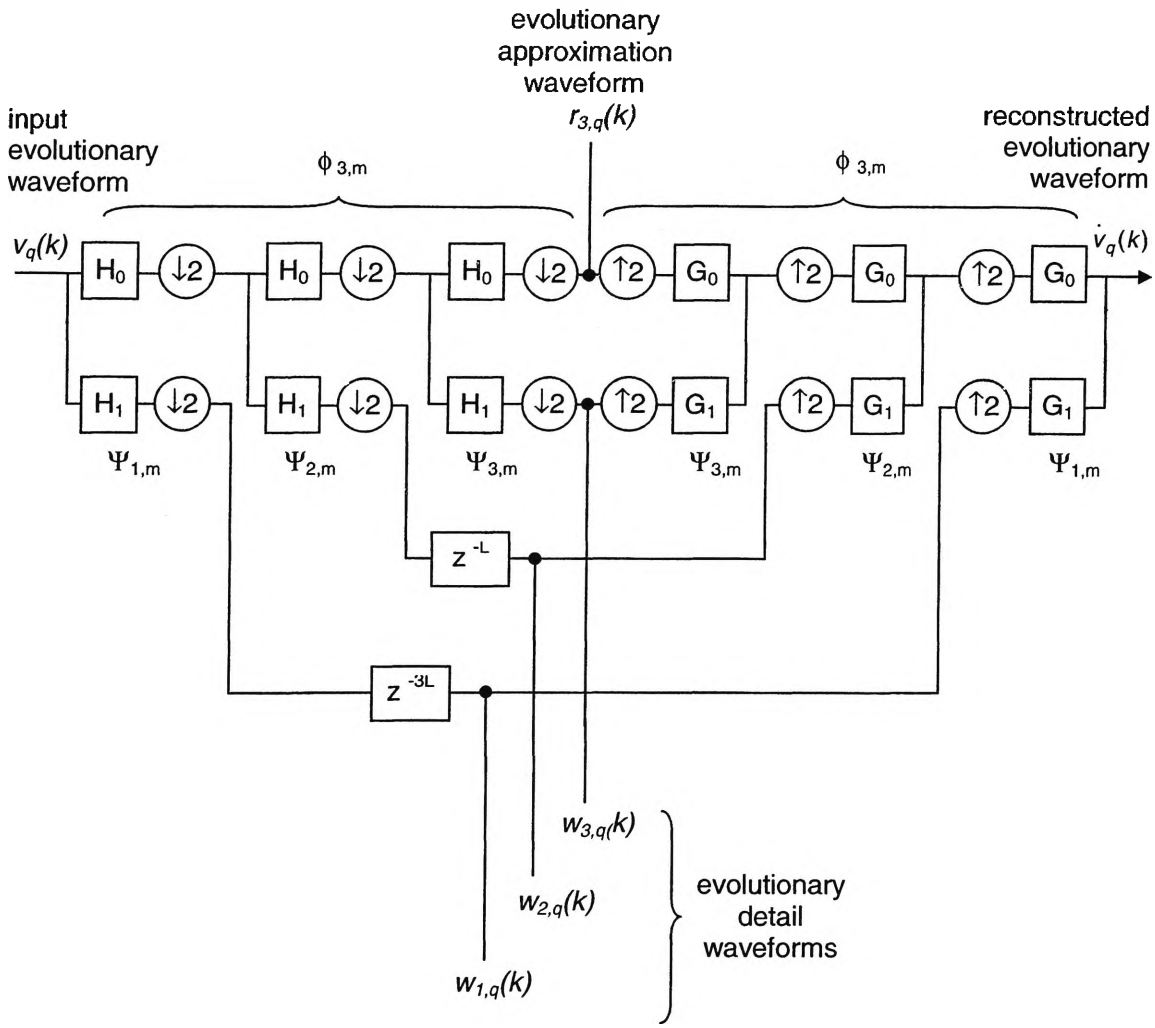


Figure 3.6. A three-level, multi-rate realisation of the wavelet decomposition and its inverse.

3.6. Wavelet Filters

The wavelet filters forming the basis of the analysis system will ideally provide good frequency resolution while allowing for perfect reconstruction. Several designs for both finite impulse response (FIR) and infinite impulse response (IIR) perfect reconstruction filter banks, which possess the necessary filter properties of causality and stability, are described here. In addition, linear phase in the passband is desirable in the FIR case. This may seem a redundant requirement, since phase distortion caused by the analysis filters is cancelled out by the synthesis filters, but in practice, non-linear quantisation errors have been found to be less detrimental if the filters have linear phase [Vaid90]. However, linear phase in the IIR case corresponds to unstable filters which are inappropriate for speech coding.

3.6.1. Biorthogonal Finite Impulse Response Wavelets

In order to obtain perfect reconstruction, i.e. no aliasing, amplitude or phase distortion, either orthogonal or biorthogonal wavelets must be used [Stran96]. The orthogonal solution has the advantage of design simplicity. However, in the two-band case, these wavelets possess non-linear phase. As a result, finite impulse response (FIR) biorthogonal wavelets derived from the Biorthogonal Spline Wavelet Family were chosen for the quadrature mirror filter (QMF) bank. Biorthogonal wavelets determined from the filters, $H_0(z)$ and $H_1(z)$, with effective lengths of 8 and 4 respectively are used, and are given by the following transfer functions:

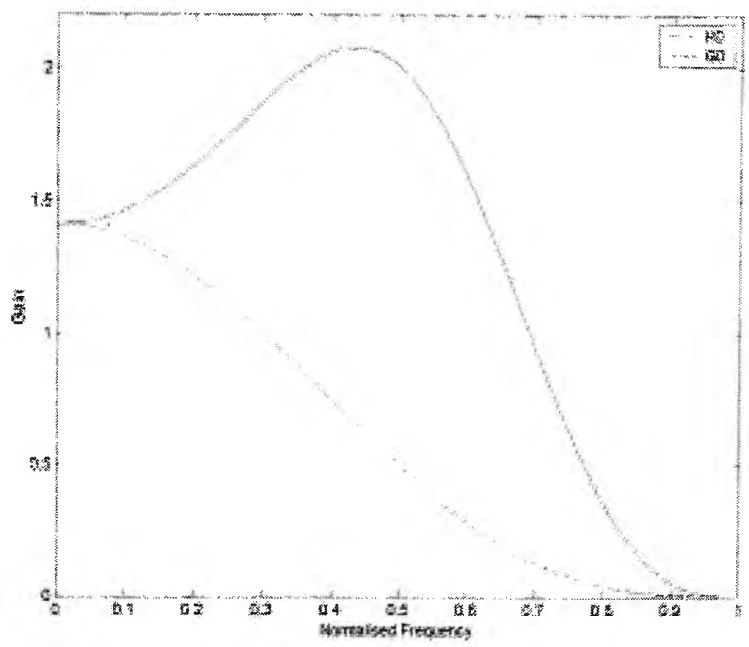


Figure 3.7. Magnitude Responses of $H_0(z)$ and $G_0(z)$ for the FIR Biorthogonal filters

$$H_0(z) = 0.0663z - 0.1989z^{-1} - 0.1547z^{-2} + 0.9944z^{-3} + 0.9944z^{-4} - 0.1547z^{-5} - 0.1989z^{-6} + 0.0663z^{-7} \quad (3.17)$$

$$H_1(z) = -0.1768z^{-2} + 0.5303z^{-3} - 0.5303z^{-4} + 0.1768z^{-5} \quad (3.18)$$

These filters possess adequate spectral characteristics while incurring a single-level analysis/synthesis delay of 7. The magnitude responses are shown in Figure 3.7. Further improvement of spectral performance (flatter passband, sharper roll-off) is at the expense of increased delay; the next highest order wavelet in the biorthogonal family incurs a system delay of 11.

While FIR filters possess the desirable properties of linear phase, and design simplicity, a significant disadvantage is the delay incurred. The overall delay of the

above filter bank for a 3-level decomposition is 49 (6 frames of 8 prototypes/frame), and increases exponentially with additional resolution levels. Since this delay is in the evolution domain, the effect of long filter delays is worsened. This inherent disadvantage of FIR filters is of major concern in real-time applications such as speech coding.

3.6.2. Infinite Impulse Response QMF Banks

One possibility for reducing the delay is to use filters with lower delay for the inner layers of the decomposition, since in a tree-structured system, the lower resolution levels located further inside the tree contribute more significantly to the overall delay than the outer levels. However, another important factor when dealing with tree-structured filter banks is the frequency separation between the bands. Poor separation can result in distorted reconstructed speech signals after quantisation, due to the inability to cancel aliased components. Both issues can be resolved with IIR QMF banks. Compared to FIR filters, IIR filters offer computational and spectral magnitude performance advantages, in addition to significant delay reductions. The progression to IIR filters gives rise to a much more complex filter design procedure, but the benefits are substantial.

Most of the literature on IIR QMF banks give solutions for filter banks possessing causal, unstable synthesis filters e.g.[Herl93][Arge96][Okud98]. These can be implemented as stable, anti-causal filters which is beneficial for image coding, however, time-reversal of the input signal is inappropriate for real-time speech coding. To satisfy the necessary conditions of causality, stability and perfect

reconstruction, the filter banks are required to be biorthogonal. Two design methods are discussed here.

Design Method 1

In this method, the filter bank is designed from a prototype filter by transformation of variables [Tay97]. Here, we consider the second simplest IIR transformation function and aim to reduce the filter overshoot. The derived analysis filter bank has the transfer functions as follows:

$$H_0(z) = \frac{-0.04 + 0.2z^{-1} + 0.59z^{-2} + 0.44z^{-3} + 0.14z^{-4} + 0.08z^{-5} + 0.03z^{-6}}{1 + 0.4z^{-2} + 0.04z^{-4}} \tag{3.19}$$

$$H_1(z) = \frac{\begin{array}{c} -0.0053 - 0.0267z^{-1} + 0.2003z^{-2} + 0.5493z^{-3} + 0.5363z^{-4} + 0.3829z^{-5} + 0.2525z^{-6} \\ + 0.1131z^{-7} + 0.0493z^{-8} + 0.0171z^{-9} + 0.0037z^{-10} + 0.0011z^{-11} \end{array}}{1 + 0.8z^{-2} + 0.24z^{-4} + 0.032z^{-6} + 0.0016z^{-8}} \tag{3.20}$$

The magnitude responses of the resulting analysis and synthesis lowpass decomposition filters are shown in Figure 3.8. These filters incur a combined group delay of 5 samples. Filter banks resulting from the design of higher order transformation functions dramatically increase group delay, and thus, they are not adopted in the current work.

Design Method 2

An alternative causal, stable IIR filter bank design procedure is that of Basu *et al.*[Basu95], in which a complete parameterisation of the solution is outlined. In this method, many intermediate parameters can be independently defined. The derived filter bank gives perfect reconstruction and incurs a delay of only 1 sample for the

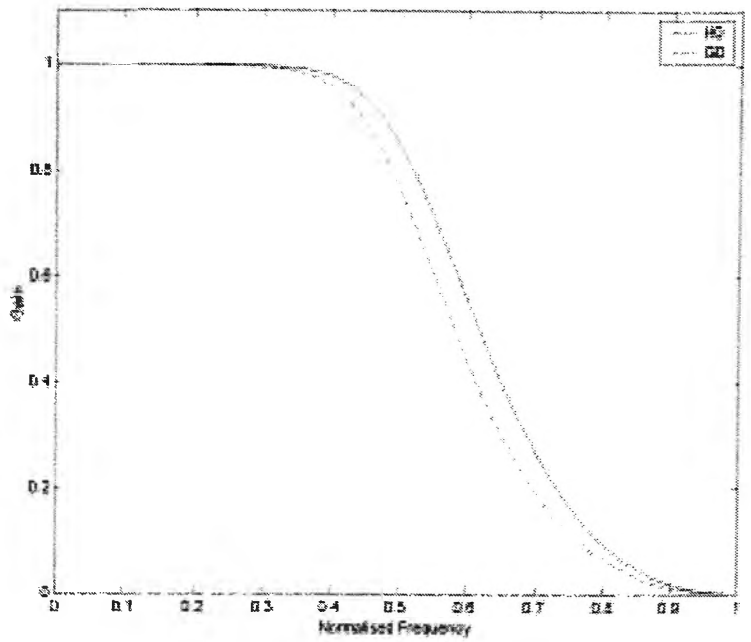


Figure 3.8. Magnitude Responses of $H_0(z)$ and $G_0(z)$ for the IIR filters of Design Method 1.

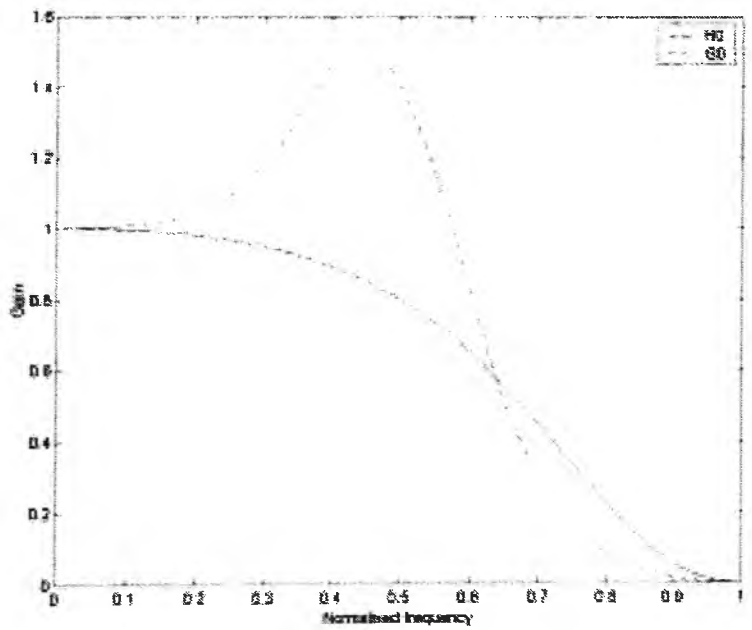


Figure 3.9. Magnitude Responses of $H_0(z)$ and $G_0(z)$ for the IIR filters of Design Method 2.

filter bank gives perfect reconstruction and incurs a delay of only 1 sample for the analysis/synthesis pair. We use a second order, lowpass half-band Bessel filter as the prototype filter, chose polynomial $r = 5z^2 + 4z + 1$, and let p and q be 2nd and 3rd order respectively. The resulting analysis filters are defined by:

$$H_0(z) = \frac{0.3987 + 0.7975z^{-1} + 0.3987z^{-2}}{1 + 0.4743z^{-1} + 0.1207z^{-2}} \tag{3.21}$$

$$H_1(z) = \frac{\begin{array}{c} 0.6897 - 1.4557z^{-1} + 0.9947z^{-2} - 0.4400z^{-3} \\ + 0.2930z^{-4} - 0.1043z^{-5} + 0.0227z^{-6} \end{array}}{1 + 0.8z^{-2} + 0.2z^{-4}} \tag{3.22}$$

The magnitude responses for the analysis and synthesis filters are displayed in Figure 3.9.

Comparison of the Filters of Design Methods 1 and 2

A comparison of Figure 3.8 and Figure 3.9 shows that the filters of design method 1 have a much more appealing frequency response with far less overshoot and much sharper roll-off. While both methods ensure perfect reconstruction of the signal in the unquantised case, the filters of method 1 produce less distortion when the magnitude and phase for each of the decomposed surfaces is quantised. Experiments showed that the filters of design method 1 tend to reduce the surface evolution bandwidth in comparison to the filters of design method 2. This allows more efficient quantisation.

On the other hand, the filters of design method 2 have the advantage of having a very low delay of only 1 sample for the analysis/synthesis pair, compared to a delay of 7 samples for a FIR filter of the same order. In a three-level decomposition, this

results in a total of only 7 samples of delay in the evolution domain (1 frame delay when extracting 8 CWs/frame).

3.6.3. Low-Delay FIR Filters

An alternative approach to address the problem of an inherently long system delay, as well as avoid the magnification of quantisation errors experienced with recursive filters, is to use low-delay FIR filters. Usually, for analysis and synthesis FIR filters of length *N*, the overall system delay is *(N-1)*. However, Nayebi *et al.* [Naye94] have presented an approach for the design of analysis-synthesis systems, in which the delay can be considered to be relatively independent of the length of the constituent filters. The system delay may be reduced by compromising filter characteristics, such as the stopband attenuation. However, imposing a very low delay on systems with relatively high order filters may not be as effective as imposing a similar delay on systems with lower order filters.

In the wavelet decomposition, the following eight-tap filters, with an imposed system delay of one sample, are used.

$$H_0(z) = 0.3896 + 0.6323z^{-1} + 0.1402z^{-2} - 0.2128z^{-3} + 0.0044z^{-4} + 0.0807z^{-5} - 0.0210z^{-6} - 0.0098z^{-7} \quad (3.23)$$

$$H_1(z) = 0.4152 - 0.6092z^{-1} + 0.1066z^{-2} + 0.1656z^{-3} - 0.0043z^{-4} - 0.0319z^{-5} + 0.0080z^{-6} + 0.0038z^{-7} \quad (3.24)$$

The frequency responses and group delays for $H_0(z)$ and $G_0(z)$ filters respectively, are given in Figure 3.10 and Figure 3.11 respectively.

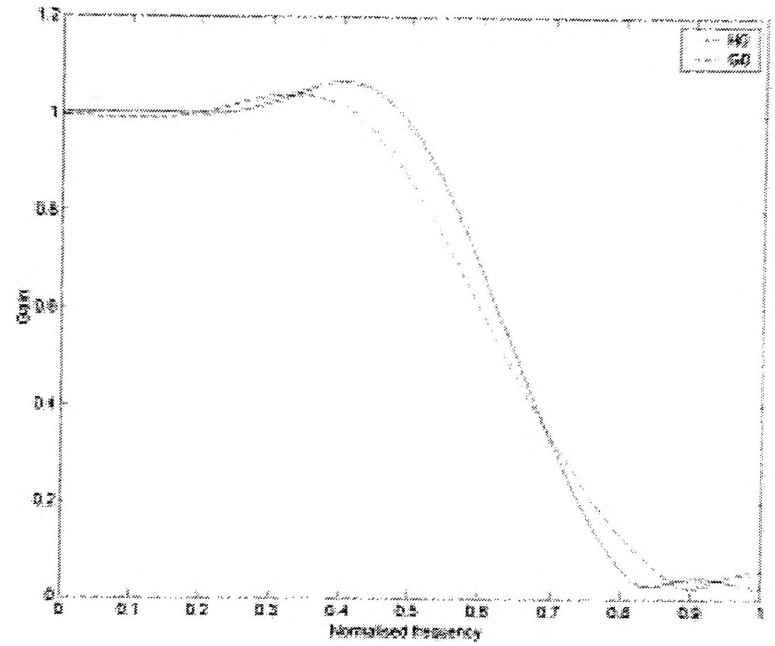


Figure 3.10. Magnitude Responses of $H_0(z)$ and $G_0(z)$ for Low-Delay FIR 8-tap filters.

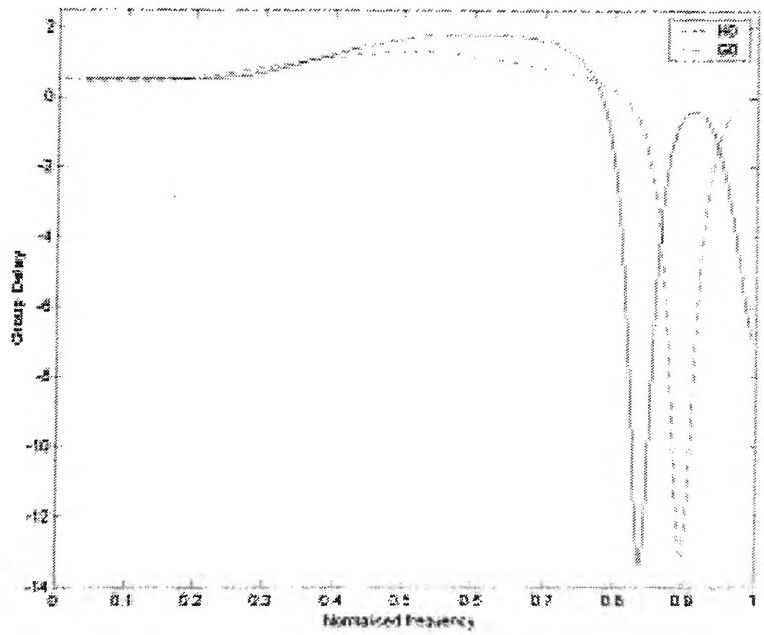


Figure 3.11. Group Delay of $H_0(z)$ and $G_0(z)$ for the Low-Delay FIR 8-tap filters.

3.7. Decomposition Of The CW Surface

A basic description of the decomposition of the CW surface was outlined in Section 2.11.2. Here, several decomposition methods for different representations of the aligned CW surface are considered. The aim is to find the method which produces surfaces that will allow the most effective quantisation.

1. Time Domain Decomposition
2. Frequency Domain Decomposition

   A: Real/Imaginary

   The evolution of the complex DFT coefficients is decomposed by filtering. This is equivalent to separating the real and imaginary components of each DFT coefficient and filtering the components individually.

   B: Separate Magnitude/Phase

   The magnitude and phase of the DFT coefficients of the CWs are separated and filtered individually. To avoid the effects of phase wrapping, each phase value is represented by a unity magnitude vector, specified by its real and imaginary components.

The surfaces obtained by the two frequency domain techniques have different characteristics. Assuming the standard WI decomposition filter, the REW component for Decomposition A will exhibit rapid changes in the combination of magnitude and phase. In contrast, Decomposition B will consider the rate of change of the magnitude and the phase as separate entities. Thus, a section of speech with

slow magnitude evolution but rapid phase evolution will dominate a detail surface in Decomposition A, but an approximation surface in Decomposition B. The question is then, which set of decomposed surfaces will allow the most effective quantisation? Analysis showed that Decomposition B produced smoother magnitude surfaces and better separation of the quasi-periodic component into the slowly-evolving surface. This is logical since both periodic and non-periodic surface sections may have identical magnitudes. The phase alone dictates the "pulse nature" of the surface.

The CW surface and its component surfaces from a 3-level wavelet decomposition of a voiced and unvoiced sound are depicted in Figure 3.12 and Figure 3.13 respectively. The filters used were low-delay FIR filters, and the Real/Imaginary decomposition was employed. During voiced speech, the signal has a quasi-periodic nature, and evolves slowly. As a result, most of the energy appears in the residue, $r_3(k,q)$, producing a smooth surface containing the underlying pulse shape evident in the CW surface. In comparison, the detail surfaces, $w_n(k,q)$ are very flat and contain only a small amount of energy. A typical decomposition of the CW energy for voiced sounds is in the ratio 1400:15:11:1 for the surfaces $r_3(k,q)$: $w_3(k,q)$: $w_2(k,q)$: $w_1(k,q)$. Although, it may seem less beneficial to have three REW-like surfaces, the advantage in multiple detail surfaces lies in the different scales of information available due to the decimation, none of which are redundant. The decomposition produces precise information about the content of each evolution frequency subband. We therefore know what information to transmit (or not to transmit) in order to achieve scalability for variable or higher rate speech coding.

The lack of energy decomposed into the highest rate detail surface $w_1(k,q)$ (Figure 3.12(a)) suggests that these coefficients may be discarded at low rates such as 2.4kbit/s. Note that these surfaces have been upsampled to the original sampling rate.

For the case of the unvoiced sounds, the CW surface is very irregular (Figure 3.13). Energy is decomposed by the wavelet decomposition to all frequency subbands, with no distinctive characteristic existing in any particular surface, as for the voiced case. A typical decomposition of the CW energy for unvoiced sounds is in the ratio 6:8:8:1 for the surfaces $r_3(k,q)$: $w_3(k,q)$: $w_2(k,q)$: $w_1(k,q)$. The resulting surfaces will be especially important for removing or enhancing certain features, for example, when the speech is corrupted by background noise.

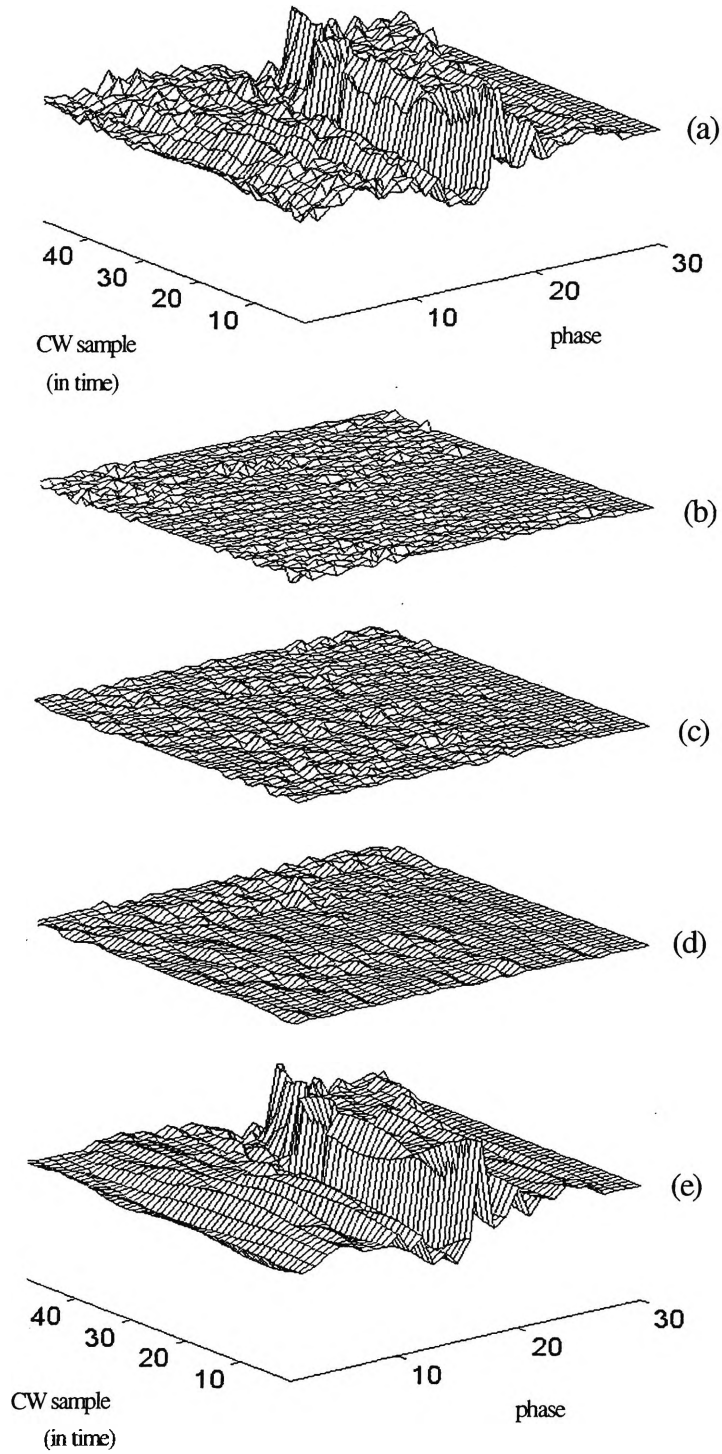


Figure 3.12. Decomposition of the voiced sound "oo" taken from the word "foolish" using a biorthogonal FIR filter bank.
(a) CW surface, (b)Detail surface, $w_1(k,q)$, (c) Detail surface, $w_2(k,q)$, (d) Detail surface, $w_3(k,q)$, (e) Approximation surface, $r_3(k,q)$

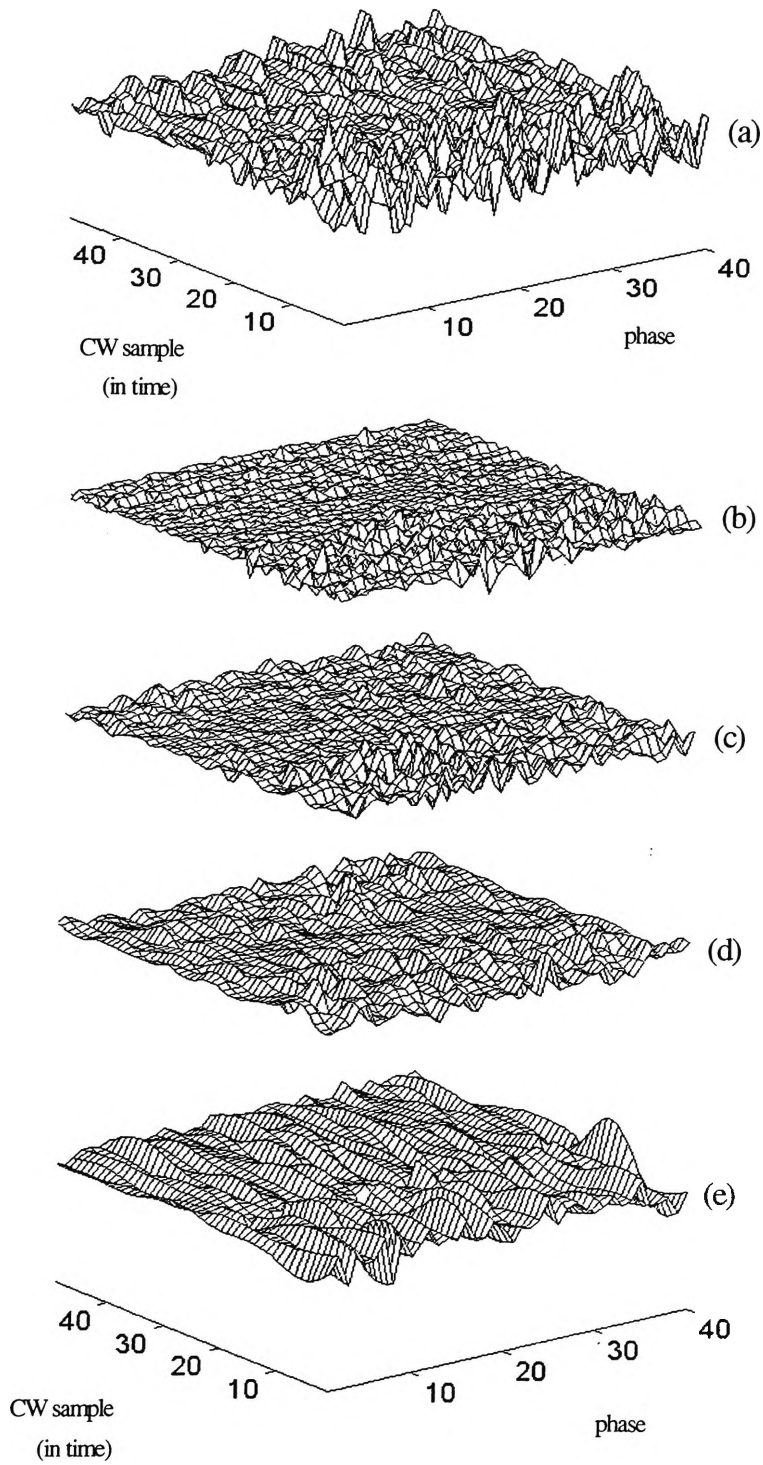


Figure 3.13. Decomposition of the unvoiced sound "sh" taken from the word "foolish" using a biorthogonal FIR filter bank
(a) CW surface, (b)Detail surface, $w_1(k,q)$, (c) Detail surface, $w_2(k,q)$, (d) Detail surface, $w_3(k,q)$, (e) Approximation surface, $r_3(k,q)$

3.8. Reconstruction of the CW Surface

The reconstruction techniques for the decomposition methods of Section 3.7 are:

1. Time domain reconstruction

   Reconstruction of the CW surfaces follows a direct reversal of the decomposition, in which the innermost approximation and detail surfaces are upsampled, passed through the synthesis mirror filters, then combined to form the approximation surface of the next highest level.

2. Frequency domain reconstruction

   A: Real/Imaginary

   The phase spectrum (either model-based or quantised) is applied to each of the individual magnitude surfaces. Reconstruction of the real and imaginary components of the DFT of the CW surface then follows a direct reversal of the decomposition. This is depicted in Figure 3.14.

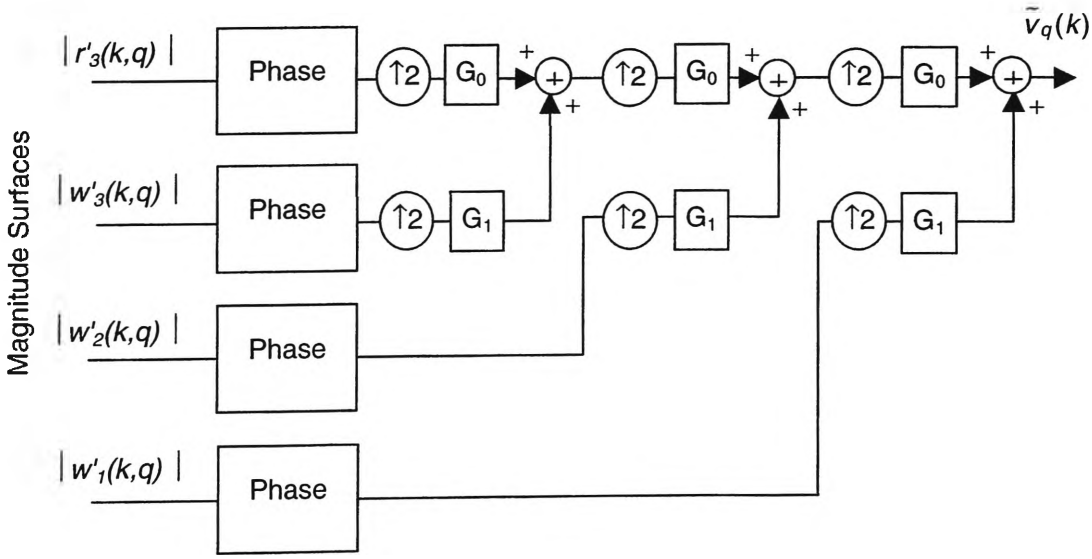


Figure 3.14. Structure of three-level wavelet reconstruction using the Frequency Domain: Real/Imaginary method.

Note that this is equivalent to individually reconstructing the surfaces up to the original sampling rate and then combining them.

B: Separate Magnitude/Phase

The magnitudes and phases of the CW surface are reconstructed separately, with each magnitude or phase surface individually reconstructed and upsampled to the original sampling frequency as shown in Figure 3.15.

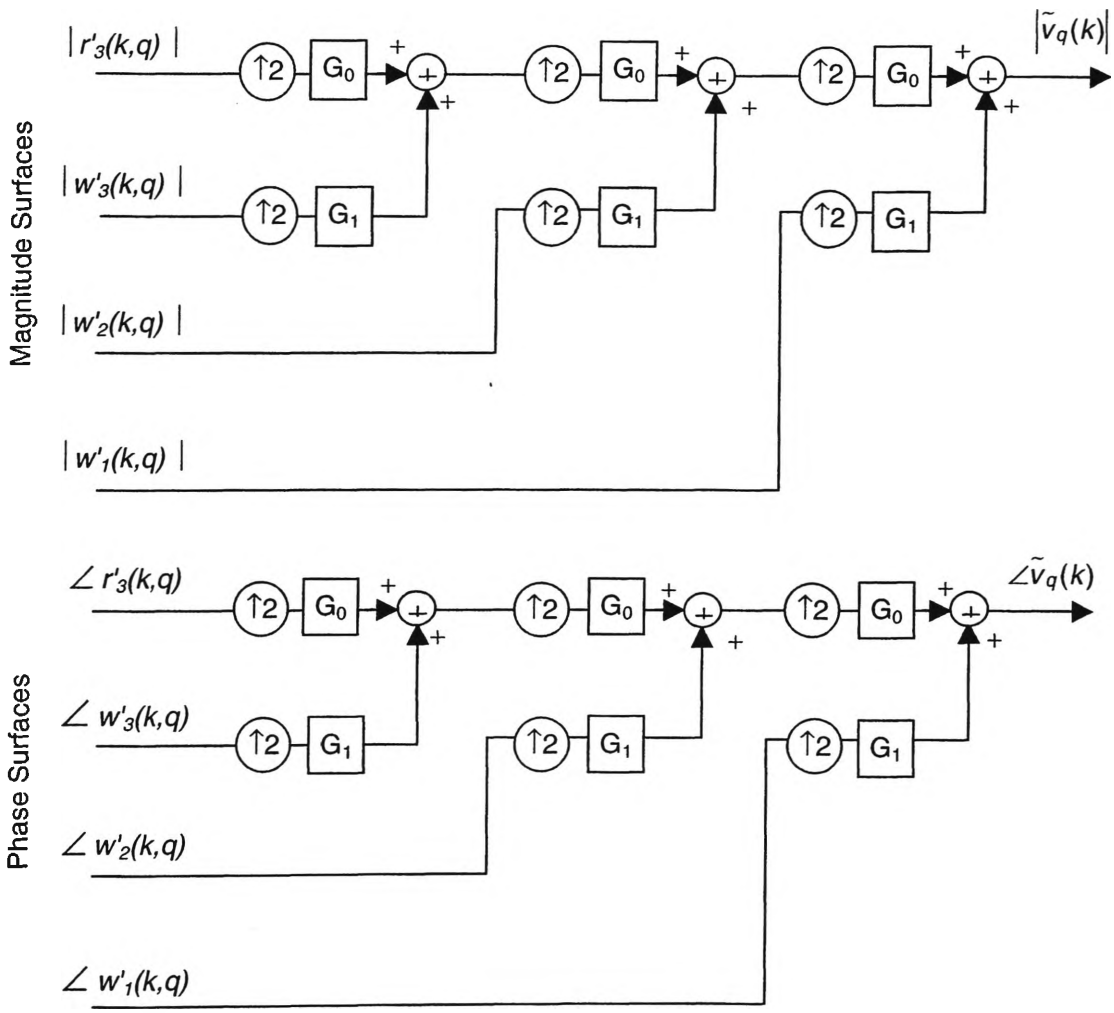


Figure 3.15. Structure of three-level wavelet reconstruction using the Frequency Domain: Separate Magnitude/Phase method, where $v_q(k)$ is the evolutionary signal, $G_0(z)$ is lowpass and $G_1(z)$ is highpass.

The choice of decomposition and reconstruction technique is largely dependent on the accuracy and effectiveness of the quantisation method used. All configurations proposed guarantee perfect reconstruction in the unquantised case, due to the mirror symmetry of the filter banks. Quantisation errors result in the non-cancellation of aliased components. These errors compound through the tree-structure of the transform, causing any errors due to the encoding of the inner surfaces to be magnified greatly. This suggests the frequency domain, reconstruction method B to be the preferred choice, as it prevents quantisation or modelling errors in the phase spectrum distorting the reconstruction of the magnitude spectrum. This issue is discussed in further detail in the following section.

3.9. Quantisation Of The Surfaces

3.9.1. Parameter Sensitivity

Evangelista [Evan93] suggests scalar quantisation schemes for the PSWT surfaces. Indeed, our results show that very coarse quantisation (2-4 bits) of individual magnitude and phase components maintains high quality output speech. This result suggests that while the exact magnitude and phase values are not significant, the *trend* of these values is important. This is particularly true with the phase parameter.

A test was performed to measure the sensitivity of the phase component to quantisation errors. Gaussian random noise was added to the phase of the surfaces

(the original magnitude of the parameters was maintained) to model the quantisation noise. The permissible level of Gaussian noise which was added to the correct phase component without noticeable audible distortion was measured. Of the three innermost surfaces ($w_{2,m}(k)$, $w_{3,m}(k)$, and $r_{3,m}(k)$), the phase of $w_{2,m}(k)$ was the least sensitive, with additive noise of variance 1.0 radians maintaining good performance. Noise with variance 0.6 radians was the maximum tolerated before distortion for $w_{3,m}(k)$, while noise of variance greater than 0.2 radians added to the phase of the $r_{3,m}(k)$ surface created distortion. Thus, the phase of the detail surfaces can be judged to be significantly less sensitive than that of the approximation, residue surface. This concurs with the phase assumptions made in the standard WI coder. In addition, the levels deeper within the tree-structure require greater accuracy than the outer levels. As explained later in Section 3.9.6, the dependence of the wavelet decomposition on phase and phase surface interrelationships adds a significant dimension to the phase quantisation/representation problem. Such dependencies are particularly relevant when considering the use of VQ schemes for surface quantisation.

3.9.2. PSWT Quantisation

In [Evan93], the time-domain waveforms of the PSWT were encoded using Adaptive PCM (APCM) for the detail surfaces, and an 8-bit uniform quantiser for the approximation surface. This required a bit rate of 21 kbit/s for the transform coefficients alone, or 9 kbit/s, when the two higher rate surfaces were omitted. Obviously, for low-rate speech compression, more efficient quantisation techniques

are desired. Here, several vector quantisation techniques, as well as standard WI quantisation methods which incorporate perceptual information, are proposed. These potentially lower the overall coder bit rate to 2.4 kbit/s. In addition, significant depth is added to the work of [Evan93] by discussing the necessary relationships between the decomposed surfaces to provide perceptually accurate reconstruction.

3.9.3. Advantages of the Wavelet Decomposition

In the same way that the SEW/REW decomposition led to increased coding efficiency, the wavelet decomposition also offers advantages. However, a significant advantage of the wavelet decomposition is that due to the decimation used, the frequency that the decomposed surfaces should be transmitted is better defined. Each surface is sent at a rate corresponding to its sampling frequency. Since the filter outputs are decimated at each level, the transmission rate required for the surfaces $w_1(k,q)$, $w_2(k,q)$, $w_3(k,q)$, and $r_3(k,q)$ is in the ratio 4:2:1:1. This contrasts the current WI decomposition method in which no optimal transmission rate for the SEW and REW is apparent to provide the most accurate reconstruction.

The wavelet decomposition exhibits potential for higher and variable rate WI coding since the bit allocation for the surfaces can be flexible, allowing a more accurate description of perceptually important scales. Also, scales that are relatively insignificant may be omitted from the reconstruction. As previously shown in Figure 3.12, the first decomposition level detail surface (the surface requiring the highest transmission rate) was seen to not add significant perceptual detail to the

speech and was thus not transmitted. In addition, the entropy of the wavelet decomposition is low, with coefficients tightly clustered about zero. This will further increase coding efficiency.

3.9.4. Application of Standard WI Quantisation Techniques

In WI, quantisation is commonly performed on the DFT coefficients. A trained codebook is used to quantise reduced-bandwidth SEW magnitude information, which has been downsampled and is transmitted once per frame. The REW magnitude spectrum is quantised using Chebyshev polynomial techniques and is transmitted multiple times per frame. For the phase component, the SEW phase spectrum is quantised using a fixed phase model derived from natural male speech and random phase is used for the REW phase spectra.

Quantisation techniques similar to those of the SEW and REW could be applied to the approximation and detail outputs of the wavelet decomposition respectively. However, these techniques need to be adapted to take advantage of the previously mentioned positive attributes of the new decomposition. In addition, increased performance can be achieved with improved matching of the Fourier coefficients, both magnitude and phase, using vector quantisation techniques

3.9.5. Perceptual Significance of the Decomposed Surfaces

To determine the perceptual importance of each of the decomposed surfaces, and hence, effectively allocate bits for quantisation, a series of pair-wise comparison tests were performed. The reconstruction configurations tested were:

- All surfaces included v Surface $w_1(k,q)$ omitted
- All surfaces included v Surface $w_2(k,q)$ omitted
- All surfaces included v Surface $w_3(k,q)$ omitted

Each sentence pair contained a sentence with one of the detail surfaces omitted from the reconstruction. Hence, in terms of error minimisation, it was easy to distinguish the sentence which produced the smallest reconstruction error. To determine the level of perceptual distortion, 25 untrained listeners listened to the sentence pairs and were asked if they preferred the first sentence, the second sentence, or could not distinguish between the two sentences, i.e. they had no preference. The results are shown in Table 3.1 to Table 3.3.

The results in Table 3.1 show that 88% of the time, test subjects either preferred speech with the first level decomposition detail surface, $w_1(k,q)$, removed above the perfectly reconstructed CW surface, or could not tell that any part of the speech had been modified. This confirms that this evolution frequency subband does not contain perceptually significant information and therefore, it is not required to be transmitted at all.

Listeners preferred speech with the second detail surface, $w_2(k,q)$, removed, or could not detect that any part of the speech had been removed 58% of the time. They also preferred speech with the third detail surface, $w_3(k,q)$, removed, or could not tell the difference between the reconstructed speech with all surfaces present and the speech with surface $w_3(k,q)$ omitted 42% of the time. These results show that the surfaces $w_2(k,q)$ and $w_3(k,q)$ contain more perceptually significant information than

surface $w_1(k,q)$ and need to be transmitted. In addition, $w_3(k,q)$ displays the most

Table 3.1. Subjective comparison results for reconstructed speech with (a) all surfaces included and (b) surface $w_1(k,q)$ removed

Preferred no surfaces removed (%)	Preferred $w_1(k,q)$ removed (%)	No preference (%)
12	14	74

Table 3.2. Subjective comparison results for reconstructed speech with (a) all surfaces included and (b) surface $w_2(k,q)$ removed

Preferred no surfaces removed (%)	Preferred $w_2(k,q)$ removed (%)	No preference (%)
42	6	52

Table 3.3. Subjective comparison results for reconstructed speech with (a) all surfaces included and (b) surface $w_3(k,q)$ removed

Preferred No surfaces Removed (%)	Preferred $w_3(k,q)$ Removed (%)	No Preference (%)
58	6	36

perceptual significance of the detail surfaces. These results form the basis for the bit allocation for the detail surfaces.

The approximation surface, $r_3(k,q)$, contains by far the most perceptually important information – the waveform periodicity and shape during voiced speech, including the majority of the signal energy. Each approximation waveform requires a significant number of bits to represent it.

3.9.6. Magnitude Quantisation

To enhance the performance of WI, vector quantisation (VQ) techniques were employed. However, the extraction of pitch-length waveforms in WI to exploit the speech periodicity, results in the need to quantise variable-length waveforms. Rather than use a multi-codebook approach, where a separate codebook is used for each possible waveform dimension (or dimension range), a Variable Dimension Vector Quantisation (VDVQ) [Das96] scheme was applied to quantise the individual surfaces. Preparation of the training data set to form the required codebooks involves the following steps:

1. A fixed codebook vector length, *L*, is chosen. This length must be greater than the length of the largest possible input vector.
2. The input vector samples/coefficients are mapped to equally-spaced positions within the fixed-length vector.
3. Zeros are then inserted in frequency bins which do not contain a value.
4. The mean of the entire training data set is calculated, then subtracted from each non-zero sample. In VDVQ, a consequence of the zero-insertion is that to correctly train the codebooks, the average value of the incoming vectors must also be zero.

For example, if the variable dimension vector is of length 3,

$$v_1 = [1 \quad 2 \quad 3] \tag{3.25}$$

and the chosen fixed codebook vector length is 6, then the training vector is

$$v_1 \rightarrow \left[\frac{1}{\dot{v}} \quad 0 \quad \frac{2}{\dot{v}} \quad 0 \quad \frac{3}{\dot{v}} \quad 0\right] \tag{3.26}$$

where v is the entire training data sequence, and $\dot{v}$ denotes the mean of v.

Codebooks are then trained on these fixed-length sequences and the resulting codebooks are mean-adjusted. During the codebook search, the codebook vectors are sub-sampled to produce a vector of pitch-length, and compared with the input vector. This technique overcomes the difficulty of quantising waveforms whose lengths vary with the pitch period.

VDVQ is used to quantise the decomposed surface magnitudes. In accordance with our perception of the different resolution surfaces, more bits are required to quantise the magnitude of the approximation surface than the magnitude of the detail surfaces.

Best quantisation results are achieved when the magnitudes are decomposed separately, independent of phase. The decomposed surfaces then have reduced dynamic range, enabling smaller codebooks to be used which offer the same quality magnitude-quantised speech as that obtained by using large codebooks with the Real/Imaginary approach. Good results were obtained using the bit allocations of Table 3.4.

Application to Standard WI Coder:

The VDVQ techniques described have also been applied to the standard WI coder for the SEW magnitude surfaces, with improved results. The technique enables efficient quantisation of complete SEW magnitudes, removing the requirement for

Decomposed Surface	Magnitude Shape (bits)	Magnitude Gain (bits)	Frequency per frame	Magnitude bits per frame
$r_{3,m}(k)$	8	3	1	11
$w_{3,m}(k)$	4	3	1	7
$w_{2,m}(k)$	2	0	2	4
$w_{1,m}(k)$	0	0	4	0
Total bits per frame				**22**
Magnitude Bit Rate (25 ms frame)				**880 bits/s**

Table 3.4. Bit allocation for shape-gain VDVQ of surface magnitudes

reduced-bandwidth spectral bin approaches. This makes the implementation of scalable and higher rate WI coders more practicable, since the low-pass nature of SEW quantisation in previous coders has been found to be an inherent limitation.

3.9.7. Phase Representation

In parametric coders, the phase is commonly not transmitted as it is of secondary importance. In these implementations, phase models are applied. However, phase sensitivity tests showed that random phase could not simply be applied to the detail surfaces within the conventional tree-structured configuration without causing considerable distortion. This can be attributed to the magnification of phase errors in the inner layers of the tree structure.

An analysis of the relationship between the phase of the original CW and decomposed waveforms was performed for the tree-structured decomposition, and it was noted that the approximation surface has similar phase characteristics to that of the original CW phase. This is particularly true in slowly evolving sections of the speech. In these highly periodic regions, the detail surfaces also retain significant phase characteristics of the original surface. Our experiments have indicated that the exact CW phase is not required to maintain good perceptual quality, but we must ensure that some degree of phase inter-relationship between the surfaces is preserved, to allow the reconstructed, upsampled surfaces to carry the correct phase characteristic trend.

To minimise phase distortion resulting in rough-sounding output speech quality, either:

1. the phase relationships between the surfaces need to be preserved to lessen the effects of compounding quantisation errors within the tree-structured transform, or
2. the reconstruction technique needs to be adapted.

To create a better description of the phase characteristic and aim towards the criteria of 1), we look at the use of VQ techniques to represent phase. Alternatively, the application of phase models in conjunction with a modified reconstruction method is also studied.

Vector Quantisation of Phase

To achieve improved phase representation than standard WI, phase quantisation, in the form of VDVQ, was applied. This has the advantage of increasing the quality without using an excessive number of bits. Some adjustments, however, were made to the VDVQ training algorithm to cope with phase wrapping. Firstly, the CW was circularly rotated so that the pulse was located at position zero. This removes the linear phase offset component, which governs the pulse position, from the phase spectrum. Absolute phase values for each DFT coefficient were then converted to a unity magnitude phase vector for training. This unity magnitude was reinforced after each codebook update. The "complex" phase vector is then used during the codebook search, eliminating problems due to the modulo 2π attribute of phase. Hence, we solve

$$\min\left|\sum_{j=0}^{\tau}\left(\left(\phi_{i,real}(j)-\phi_{c,real}(k)\right)^2+\left(\phi_{i,imag}(j)-\phi_{c,imag}(k)\right)^2\right)\right| \tag{3.27}$$

where,

$$\begin{aligned}\phi_{real} &= \cos\phi,\\ \phi_{imag} &= \sin\phi,\\ k &= \frac{jL}{\tau},\end{aligned} \tag{3.28}$$

and ϕ_i and ϕ_c denote the input and codebook phase values, τ is the pitch period value, and L is the fixed codebook size.

The phase vector selection was constrained during voiced sections with the aim of maintaining a level of phase inter-relationship during these perceptually important periods. This is done by using sectioned codebooks, and training voiced and

unvoiced phase spectra independently. The phase codevector for the approximation surface is first chosen, then limits are placed on the phase vector selection for the detail surfaces, rather than quantising the phases of all surfaces independently. As an alternative approach, the phase of the CW surface may be quantised. To keep within the perfect reconstruction framework, however, this results in multiple transmissions of the phase, and fails to take advantage of the lower-rate surfaces of the wavelet decomposition and human perception of phase.

The effects of low-rate quantisation of the surfaces are more significant for IIR filters than for FIR filters due to their recursive nature, causing a swirling distortion or reverberation if inter-surface phase relationships are not adequately maintained. Hence, of the described filter designs of Section 3.6, low-delay FIR filters are preferred due to their short group delay and less complicated quantisation task than the IIR filters.

The low-order decomposition filters required to keep the delay minimal, result in a high level of aliasing, which is difficult to cancel at low rates. Thus, while phase is of secondary importance in speech coding, any errors due to the encoding of the inner surfaces within the tree-structure are magnified greatly. This makes vector quantisation of phase not as effective within the tree-structured wavelet decomposition, as it may be for quantisation of the SEW and REW phase components of standard WI as described earlier. The analysis-by-synthesis phase quantisation approach of [Gott99][Gott00] is also impractical here, due to the large

number of additional operations required to synthesise the vector candidates up the tree structure.

Phase Modelling

The second technique for good phase representation involves adapting the reconstruction method. Instead of reconstructing the combined magnitude and phase contributions for each surface, independent reconstruction of these components removes the interference of phase quantisation errors from the magnitude reconstruction.

It was initially proposed in [Chon99] that the surface magnitudes could be individually reconstructed up to the original sampling frequency, then a phase model for each surface could be applied. The incorporation of phase was performed prior to the recombination of the component surfaces to form the CW surface, as shown in Figure 3.16. The rationale for this technique lay in the desire to apply a modelled phase to one surface without affecting the phase of higher rate (outer) surfaces, which is the disadvantage of the structure in Figure 3.14. The implementation, however, is flawed, in that, in order to accurately reconstruct the CW magnitudes, the component surface magnitudes must be directly added (with all frequencies in phase). It is also noted, and depicted in the time-domain waveforms of Figure 3.18, that the filter bank produces an unusual modulation effect on the upsampled surfaces, particularly noticeable in the reconstructed $w_3(k,q)$ waveform. Hence, while the Separate Magnitude/Phase decomposition and reconstruction is still advantageous to allow accurate magnitude reconstruction, the

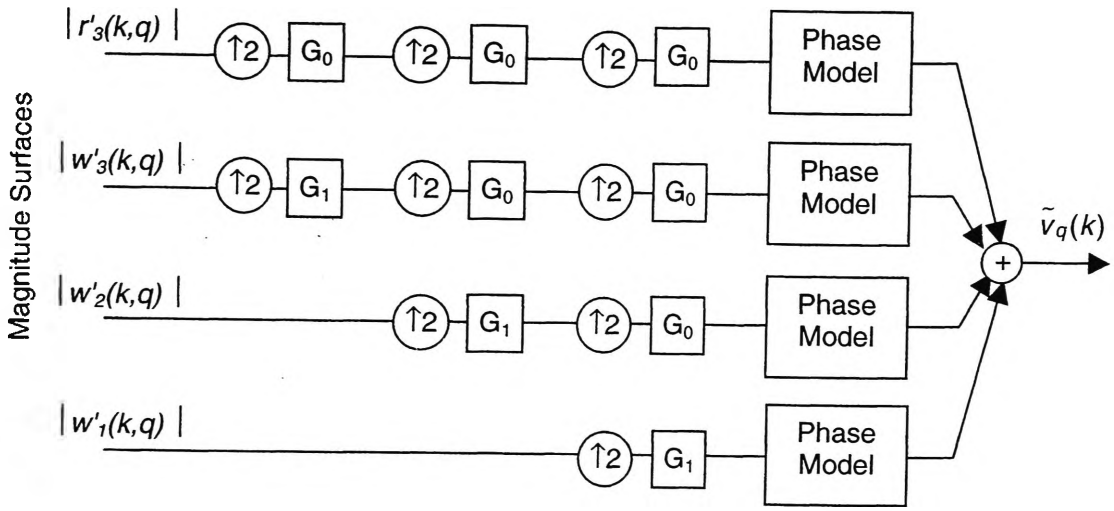


Figure 3.16. Structure of three-level wavelet reconstruction using the Frequency Domain: Separate Magnitude/Phase method with multiple phase models, where $v_q(k)$ is the evolutionary signal, $G_0(z)$ is lowpass and $G_1(z)$ is highpass. The modelled phase for the each of the surfaces is applied to the upsampled surface magnitudes

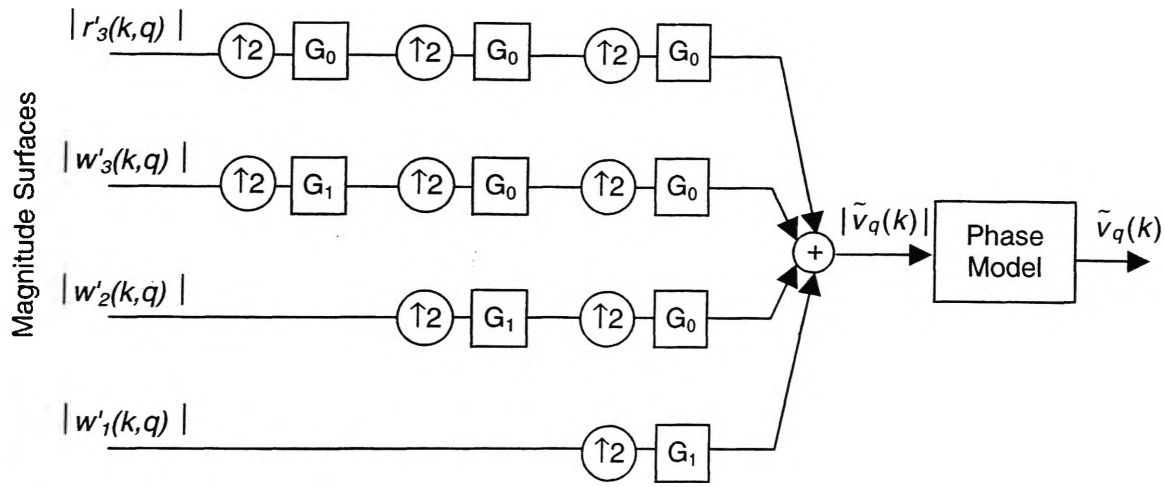


Figure 3.17. Structure of three-level wavelet reconstruction using the Frequency Domain: Separate Magnitude/Phase method with a combined phase model, where $v_q(k)$ is the evolutionary signal, $G_0(z)$ is lowpass and $G_1(z)$ is highpass. The phase for the complete signal is applied to the combined surface magnitudes.

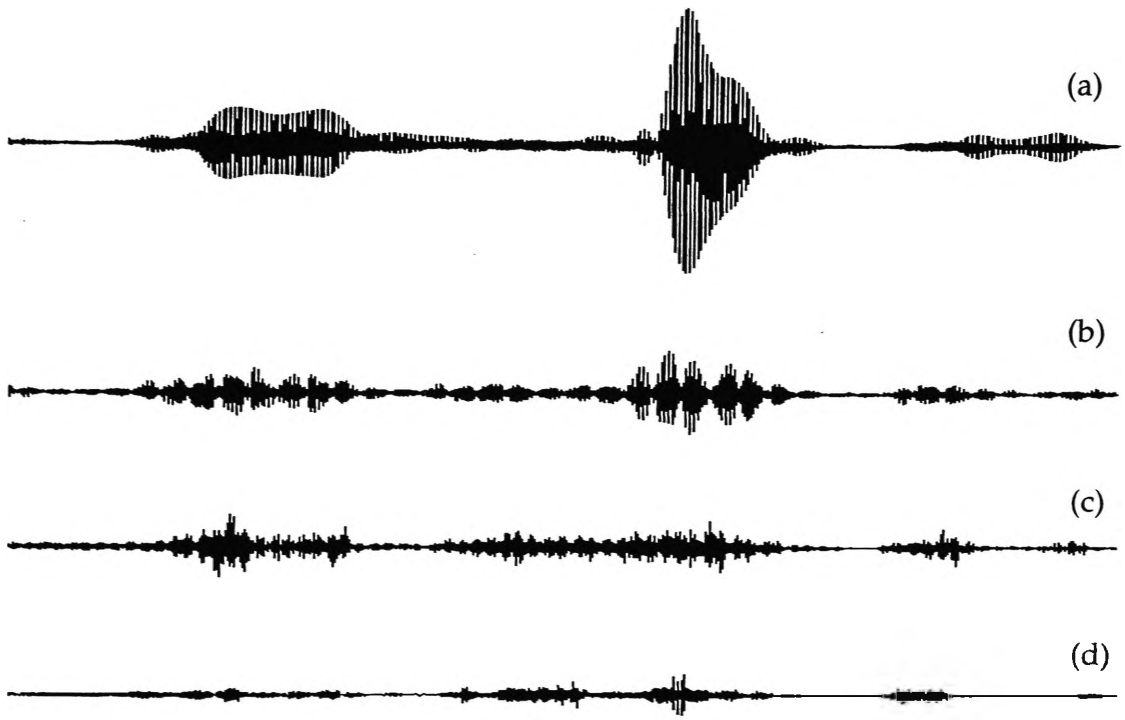


Figure 3.18. Individual reconstruction of each wavelet-decomposed surface up to the original sampling frequency. (a) $r_3(k,q)$, (b) $w_3(k,q)$, (c) $w_2(k,q)$, (d) $w_1(k,q)$.

phase must be applied to the combined surface magnitudes, rather than to each surface individually. The corrected structure is depicted in Figure 3.17. This requires representation of the CW phase.

In existing low-rate coders, e.g. WI, sinusoidal coders, LPC vocoders, phase models are used; some requiring a voiced/unvoiced decision. The phase model utilised in this approach should ideally take advantage of the available information regarding the different rates of evolution of the CW surface. Thus, the CW phase model is based on the energy content of the decomposed surfaces. The ratio of surface energies essentially represents the voiced/unvoiced content in the speech, and hence can be used as a measure of the level of disruption of the phase linearity. This

corresponds to a dynamic voiced/unvoiced mix measurement. Since quantisation may alter the surface energies, a good indication of the "voicedness" of the speech which is available at the decoder is given by the ratio $\frac{Energy(r_{3,m}(k))}{Energy(w_{3,m}(k))}$. The gains of these two surfaces are quantised and transmitted every frame, as shown in Table 3.4. Knowledge of the phase distribution for each of the decomposed surfaces is also desired in order to model the phase well. The REW surface of standard WI is assumed to be normally distributed in the time domain, which corresponds to a uniform phase distribution. Analysis of the decomposed surface phases, and the individually reconstructed surface phases shows that the phase distribution of all the detail surfaces is also uniform.

There are several considerations in allocating phase to the CW surface:

1. The degree of linearity/randomness of the phase,
2. The contributions from each evolution frequency subband, and
3. The relative perceptual significance of the phase component

The phase model used consists of a base linear phase, with randomness (uniform distribution) added to a proportion of the coefficients, depending of the energies in the approximation and detail surfaces (at the downsampled rate). The rate of phase evolution is limited, in slowly evolving sections, by allowing the phase of the current waveform to fluctuate by a limited amount from the phase of the previous waveform. Finally, the phase spectrum is weighted by the inverse of the LPC power spectrum, to ensure that formant frequencies are not distorted.

3.9.8. Time-Domain Quantisation

VDVQ can also be used to encode the time-domain decomposed waveforms. However, difficulty arises since phase information is still contained in the surfaces, and is not able to be controlled explicitly. This prevents the use of well-known phase approximations.

3.10. Preferred Decomposition/Reconstruction Structure

Overall, for best results, the magnitude of the CW surface should be decomposed and reconstructed separately, then combined with the modelled phase representation, as depicted in Figure 3.17. Using this method, magnitude codebooks can be small, with the bit allocations of Table 3.4, and a phase model can be applied, requiring no extra bits to be sent. By choosing a good phase model which exploits the waveform evolutionary characteristics, as described earlier in this section, good quality output speech can be achieved. Phase representation could be further improved by employing more sophisticated phase models, similar to those developed for sinusoidal transform coders, such as those which use minimum phase assumptions [McAu95] or allpass filtering techniques [Ahma98].

3.11. Complexity

The increased facility due to the further separation of the speech evolution by the wavelet decomposition is at the expense of computational complexity. For the decomposition of DFT magnitudes, using 8 CWs/frame and a filter order of 7, the

number of multiplications and additions required per frame is outlined in Table 3.5. The number of computations is variable, depending of the pitch period, τ, of the particular frame of speech. The reconstruction requires twice the number of multiply-and-add operations as the decomposition, due to upsampling.

	Multiplications/frame
1st decomposition level	$4 \times 7 \times \tau/2 \times 2$
2nd decomposition level	$2 \times 7 \times \tau/2 \times 2$
3rd decomposition level	$1 \times 7 \times \tau/2 \times 2$
TOTAL	**$91 \times \tau/2$**

Table 3.5. Multiplication and addition operations required for a 3-level decomposition

We acknowledge that the scheme being proposed is complex, though the improved multi-scale analysis provides advantageous signal analysis, and enhanced coding of speech in the presence of background noise. It is also suitable for high quality speech storage applications.

3.12. Summary

The wavelet decomposition offers an alternative description of signal evolution which can be easily applied to the WI framework. Its multi-resolution analysis enables further characterisation of the voiced and unvoiced speech components.

In a similar fashion to the SEW/REW decomposition, the proposed decomposition technique operates along the evolution of pitch-cycle waveforms, exploiting the

quasi-periodicity of voiced speech. The DWT is implemented as a quadrature mirror filter bank, and several causal, stable, perfect reconstruction filter bank designs are studied. FIR wavelet filters are impractical for use in real-time applications due to the substantial delay incurred. This can be overcome by using causal, stable IIR QMF banks which incur reduced delay and produce surfaces with smooth characteristics for efficient quantisation. In addition, the improved roll-off characteristic of the IIR filters results in a more defined separation of the periodic and aperiodic fluctuations by faster tracking of the dynamic aspects of the evolutionary surfaces. Quantisation of the decomposed surfaces, however, is a challenging issue, and is worsened by the recursive nature of the filters. This led to the application of low-delay FIR filters.

The separation of the signal into evolutionary frequency subbands enables flexible bit allocation. Subjective tests were used to estimate the perceptual significance of each of the decomposed surfaces, and hence decide on the required codebook size for the surface magnitudes. Since the length of each magnitude vector varies with the fundamental frequency of the speech, VDVQ was applied. Complications associated with the quantisation of the decomposed surfaces have been identified. In particular, phase relationships between the surfaces need to be substantially maintained, and this cannot be achieved sufficiently using direct quantisation or modelling techniques. The key to maintaining high quality quantised speech lies in the ability to separate magnitude and phase in the decomposition and reconstruction. This prevents the interference of phase errors from distorting the magnitude reconstruction. In addition, the compounding of phase quantisation

errors within the tree-structure can be disastrous, leading to perceptually annoying artefacts, and hence, it is much more desirable to apply phase to the final upsampled surface. The phase model applied takes advantage of well-established perceptual qualities of voiced and unvoiced sounds, by using characteristics of the speech evolution.

The wavelet decomposition and quantisation techniques discussed in this chapter may also be applied to other transforms, such as those operating pitch-synchronously in [Yang98].

Chapter 4

Waveform-Matched Waveform Interpolation Coding – Analysis Techniques

"All exact science is dominated by the idea of approximation."

*-- **Bertrand Russell***

4.1. Introduction

The quality of waveform coders, such as the Code-Excited Linear Predictive (CELP) coder, degrades rapidly at rates below 4kbit/s. This is due to the inability of these coders to adequately represent the waveform details with the number of bits available. Conversely, parametric coders, such as the Waveform Interpolation (WI) coder, produce good quality speech at low bit rates, but their performance is limited at higher rates by the speech production model. Hence, to achieve toll-quality speech at 4kbit/s, it seems advantageous to combine the favourable attributes of both these coders - the waveform matching properties of CELP, and the effective decomposition and quantisation techniques of WI. Indeed, original WI

proposals were multi-modal WI/CELP implementations [Klei93b][Burn93], and more recently harmonic/waveform[Shlo98] and MELP/CELP [Stac00] combinations have been proposed. These are aimed towards improved coding of voiced, unvoiced and transitional regions by switching between algorithms. In this work, waveform coding and parametric coding attributes are integrated into a single coder framework; no switch is required.

A technique incorporating the waveform coding objective in the WI coder was first described by Yang and Kleijn [Yang98][Klei98]. In this method, the input speech is time-warped to have a constant pitch period, allowing the application of time-invariant parameter extraction methods which provide for exact reconstruction. A critically-sampled perfect reconstruction filter bank is then used to obtain a representation of the signal suitable for quantisation. The Waveform-Matched Waveform Interpolation (WMWI) coder proposed in this thesis builds upon the work of [Klei98], shifting the emphasis from the multi-rate filter bank approach, to achieve the primary objectives of

a) providing an accurate means of signal analysis,

b) enabling efficient perception-based quantisation techniques,

c) achieving time-synchrony and waveform matching, and

d) improving scalability to higher rates.

In this chapter, the analysis techniques of WMWI are described. These include the construction of an optimal pitch contour, pitch-normalisation of the residual signal, formation of an accurate representation of signal evolution to enable effective

decomposition, and various decomposition mechanisms to provide for efficient quantisation. Chapter 5 will build on this basis, covering methods to quantise the decomposed components and accurately transmit the pitch parameter to allow waveform matching.

4.2. Overview of WMWI Analysis

In common WI, CWs are extracted at regular intervals and aligned via cyclic rotation, as described in Section 3.3. However, the alignment procedure destroys the relative phase information of the waveforms, preventing the input signal from being recovered exactly. Similarly, speech information is lost in the asymptotically perfect reconstruction WI technique recently proposed by Eriksson [Erik99]. In [Erik99], pitch periods are selectively extracted from the residual signal to form a modified input signal, and consequently, samples of the original signal may be selected in more than one pitch period, or alternatively, not at all. These small errors accumulate, necessitating a large proportion of a pitch period to occasionally be repeated or skipped in order to retain time-synchrony with the input.

The WMWI approach addresses the issue of unrecoverable signal information, preserving all input samples during the analysis process. This produces a CW surface with an improved representation of signal evolution and enables exact reconstruction. The analysis operations, shown in Figure 4.1, consist of time-domain warping the LP residual to have a constant pitch, extraction of consecutive pitch periods to form the CW surface, and decomposition of the surface to produce

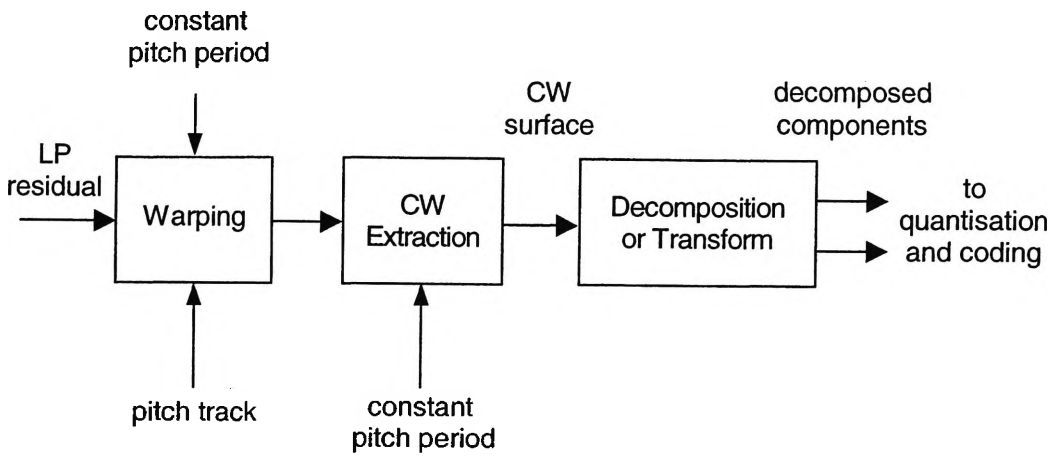


Figure 4.1. Block diagram of the significant WMWI analysis operations

components which are more convenient for quantisation. Critical-sampling of the CWs results in each sample being represented once, and only once, in the CW surface.

The most crucial of the analysis procedures is the warping operation. If the pitch track specifying the degree of warping is correct, alignment of the CWs in the generated CW surface is inherent; no further alignment routine is necessary. This allows the critically-sampled pitch periods to be effectively transformed (using DFT, MLT techniques etc. as detailed later in Section 4.5) or decomposed in the evolution domain (Section 4.6), facilitating efficient quantisation. The task of perfect alignment is not an easy one to achieve. Poor alignment of the significant features of the pitch periods not only causes a failure to adequately take advantage of the pitch periodicity of the speech signal, but results in ineffective application of VQ for the decomposed components. This indicates the need to optimise the pitch track.

4.3. Time-Domain Warping

Time-domain warping removes the pitch variations from the input signal, normalising the pitch to a constant value. This allows the use of pitch-synchronous, fixed window length parameter extraction methods, which are advantageous for signal encoding. Modification of the signal to have a constant pitch also avoids errors due to the misinterpretation of signals with rapidly time-varying pitch for noise-like, aperiodic signals. A block diagram of the warping operation is depicted in Figure 4.2. Linear prediction analysis is performed prior to warping to separate the formant structure from the quasi-periodic excitation or residual signal. This allows pitch pulses to be more easily identified, and the independent coding of the spectral shape and vocal tract excitation.

Warping involves creating an invertible mapping which associates points on the original time-scale, with points on the warped time-scale. The input speech signal, $x(t)$, has a quasi-periodic nature and thus can be modeled as a Fourier series with time-varying amplitudes, $a_i(t)$ and time- varying fundamental frequency, $\frac{1}{p(t)}$,

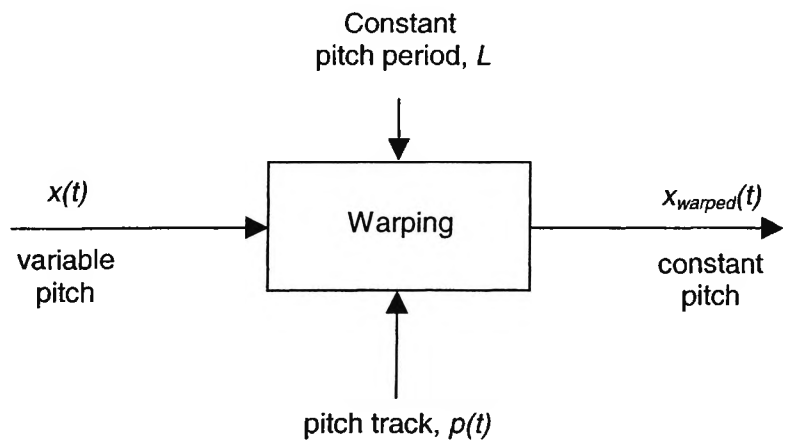


Figure 4.2. The warping operation

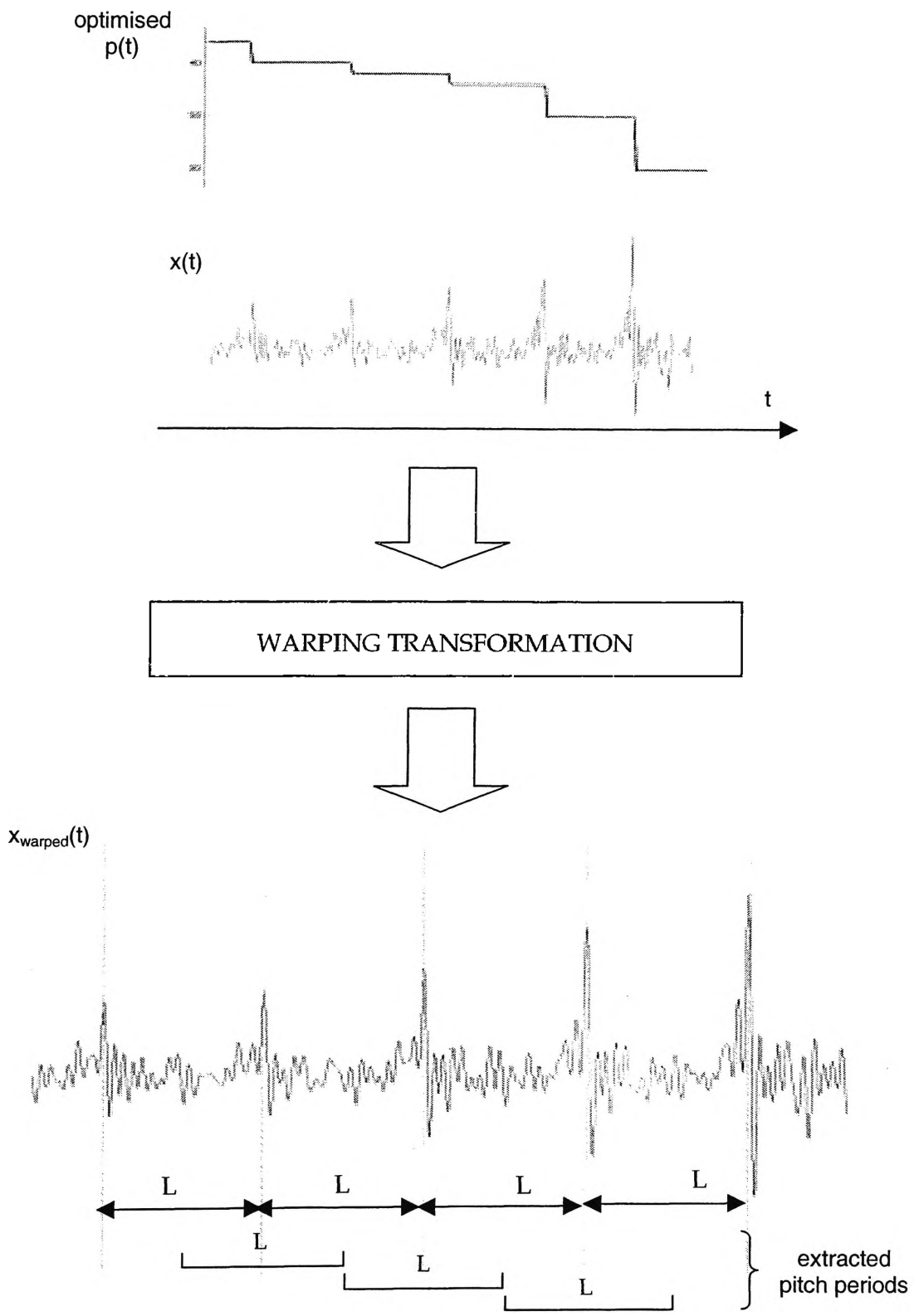


Figure 4.3. Time-domain warping of a speech signal to have a constant pitch

where $p(t)$ is the pitch track.

$$x(t) = \sum_i a_i(t) e^{ji\phi(t)} \tag{4.1}$$

where

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} \frac{2\pi}{p(\tau)} d\tau . \tag{4.2}$$

The phase track, $\phi(t)$, is a monotonically increasing, continuous function, and therefore its inverse, $\phi^{-1}(t)$, exists. Time-warping of a signal to have a constant pitch can be expressed by

$$x_{warped}(t') = x\left(\phi^{-1}\left(\frac{2\pi t}{P}\right)\right) \tag{4.3}$$

where the P is the enforced, constant pitch period length in seconds.

The mapping of a quasi-periodic signal, $x(t)$, to the warped time-domain is depicted in Figure 4.3. The interpolation filter used for the warping transformation is a hamming-windowed sinc function, and the associated optimised pitch track, $p(t)$, is derived using the methods described later in Section 4.4.

4.3.1. Warping Requirements

Accurate time-warping is crucial for good performance of the paradigm, and this is consistently emphasised in [Klei98]. However a technique to achieve this accuracy was not described. The effects of incorrect warping, such that pitch period lengths are not exactly equal, create difficulties for subsequent operations. If transforms are to be applied to the warped signal, pitch normalisation errors result in significant aliasing effects. If the SEW/REW decomposition of WI, as previously described in

Section 2.11.2, is to be implemented, effective decomposition relies on the extracted pitch periods being well-aligned. This corresponds to correctly warping the residual signal.

The objective of warping a signal to have a fixed period is difficult to achieve precisely. However, for the purpose of facilitating a useful transformation and/or decomposition of the CW surface, it is more important to align the pitch pulses than attempt to accurately warp each individual sample of the input signal. The consequences of poor pulse alignment render VQ schemes inefficient, as illustrated in Section 4.3.3. Hence, the necessary criteria for the optimal pitch track are motivated by the requirements for effective decomposition and quantisation.

Warping the residual signal with an optimal pitch track will produce a CW surface in which:

a) the length of the CWs is constant,

b) consecutive pitch pulse peaks are aligned,

c) pitch pulse peaks following an unvoiced region are also aligned with those pulses peaks preceding that section, and

d) each CW has low energy at its boundaries.

To achieve the above objectives, the pitch track is designed to warp pulse peaks to a fixed position within each warped period. This position is chosen to be the central sample of the pitch period to minimise discontinuities at the period boundaries. It is emphasised that the true pitch contour, which reflects the nature in which the glottis opens and closes during speech production, may not be the optimum pitch track for

good signal analysis and decomposition. Hence, as in methods such as Relaxed CELP (RCELP) [Klei93], the aim of the method for selecting the pitch contour is not to precisely model the rate of vocal chord vibration. While RCELP aims to achieve coding efficiency of the pitch parameter without resulting in perceptual degradation, the objective of the pitch track creation for WMWI is to provide for perfect reconstruction of the original, unmodified signal, as well as to gain coding efficiency for the quantisation of the excitation signal.

4.3.2. Definition of Terms

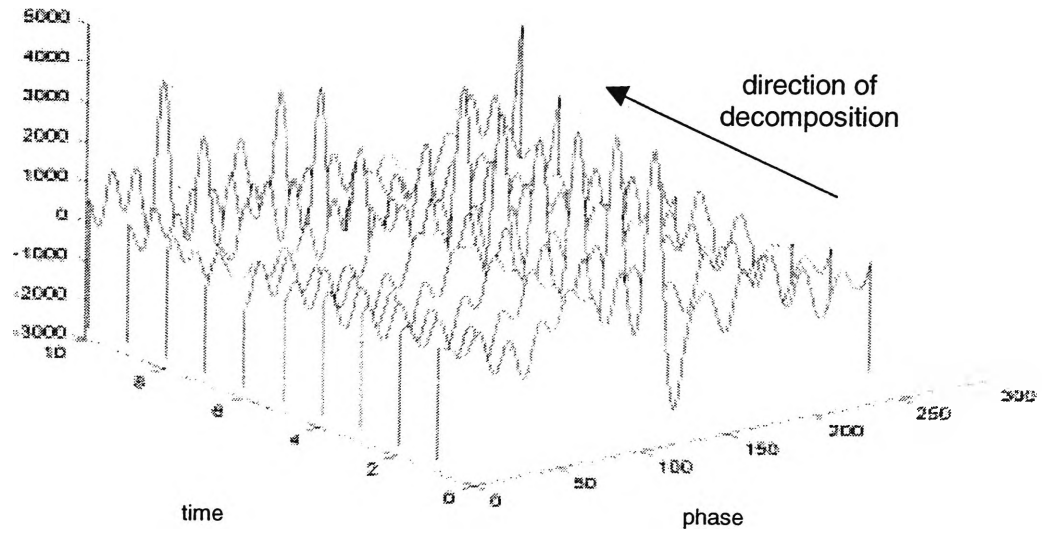
For the purpose of correctly warping to align pitch periods, the following terms are interpreted as follows:

- Frames which contain sections of high periodicity and exhibit clear pulse peaks in the residual signal are labelled as *voiced* or *pulsed*, otherwise they are *unvoiced* or *unpulsed*.
- The *pitch period*, during pulsed frames, is the distance (in samples) between adjacent pulse peaks. Hence, every cycle has an associated pitch. During unpulsed frames, the pitch has no clear definition – it is simply assigned a predetermined value, to allow continuous time-warping.

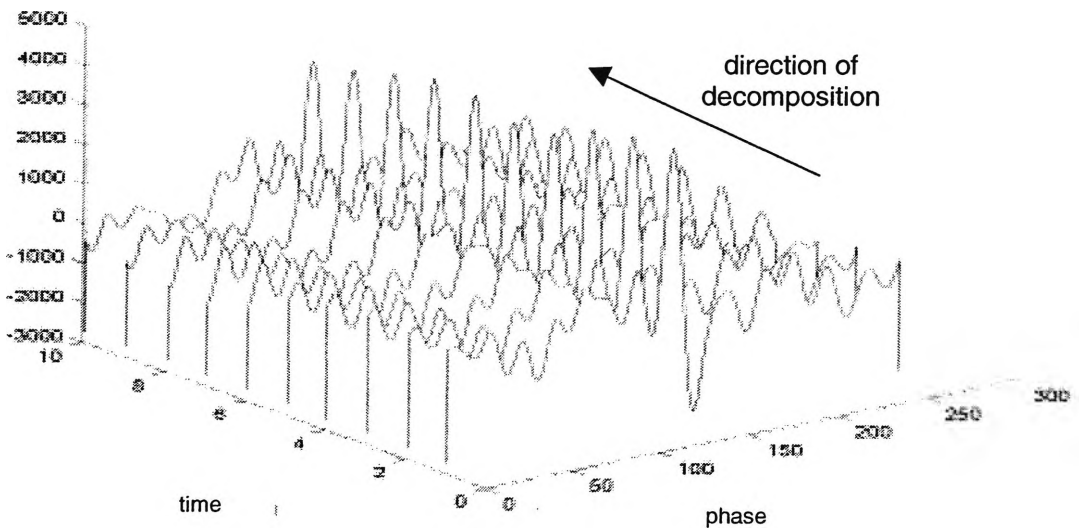
4.3.3. Effect of a Non-Optimal Pitch Track

Time Domain

The effect of warping with an incorrect and correct pitch track for a section of voiced speech residual is shown in Figure 4.4. Let us assume a decomposition similar to common WI, in which the CW surface evolution is filtered to separate the slowly-



(a)



(b)

Figure 4.4. The effect of poor pitch tracking. Characteristic waveform surface formed by extracting consecutive periods from the warped residual in the case where a) the pitch track is not correct, and b) the pitch track is correct

evolving and rapidly-evolving components, is applied. If the pitch track is non-optimal (Figure 4.4a), the non-alignment of pitch periods will cause quasi-periodic pulses of the residual signal to be decomposed into the REW. This makes REW quantisation difficult, forcing the quantisers to cope not only with noisy variations, but also with pulses with concentrated energy. The well-aligned pitch pulses of Figure 4.4b will lead to most of the signal energy being separated into the SEW, as desired. Hence, to achieve efficient quantisation, it is imperative for the pitch track to be optimised, in the sense that the possibility of pitch pulse peaks being incorrectly decomposed into the REW is minimised. Even if no decomposition is used, the lack of CW alignment makes VQ techniques ineffective as there is no strong commonality between the waveforms.

Transform Domain

The consequence of poor pitch tracking is also detrimental to the decomposition of evolving transform coefficients. If fixed-length transforms are applied to the warped CWs, it is important that similar features occur at the same time location within each CW for best interpretation and comparison of the coefficient series. Real-coefficient transforms, especially, are very sensitive to the position of the pulse within the period as the magnitude and phase characteristics are not separated.

The effect of a circular rotation of the warped period on its corresponding transform coefficients was examined to illustrate the consequence of pulse non-alignment. An experiment was performed whereby a sinc waveform of length 256 samples was transformed multiple times. For each transform, the time-domain waveform was

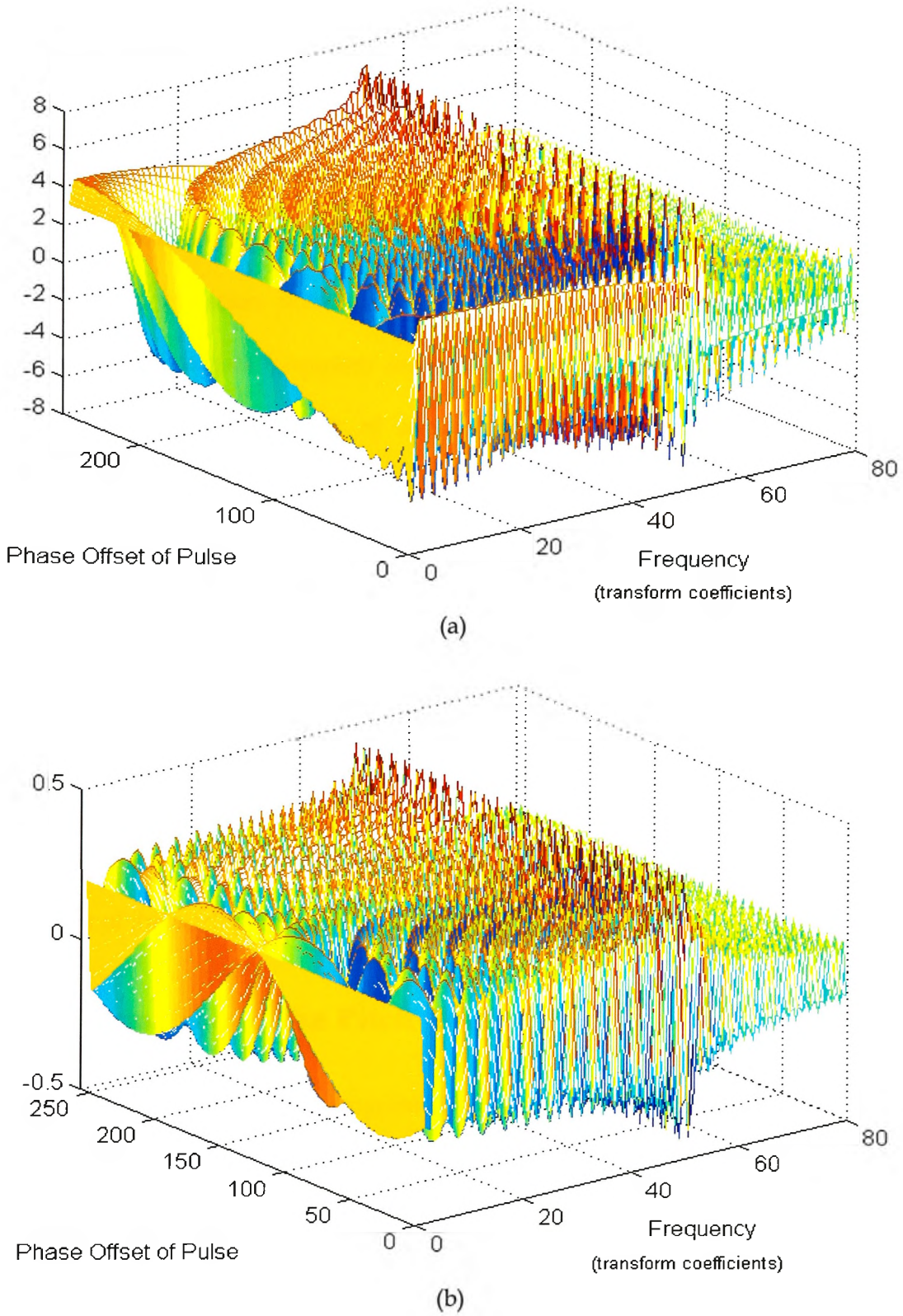


Figure 4.5. The effect of circular rotation of a pulse period (i.e. changing the phase offset of a pulse) on the transform coefficients for the a) Discrete Cosine Transform b) Modulated Lapped Transform

circularly rotated by one sample, until a full rotation was completed. The results showing the Discrete Cosine Transform (DCT) and Modulated Lapped Transform (MLT) coefficients as a function of the position of the pulse peak within the waveform (phase offset) are illustrated in Figure 4.5. The main area of activity (high energy, rapid variation within a particular coefficient) is during the first 50 coefficients, with higher frequency coefficients decaying with increasing frequency. In both transforms, a pattern of arcs is formed in the surface of transform coefficients as the phase offset is varied, the effect occurring at twice the rate for the MLT, due to the transform overlap. It can be seen that periods containing phase offsets of around ±5 samples do not produce vastly different transform results to the non-offset equivalent, but it is clear that the number of peaks in the coefficient series, as well as the overall envelope, vary dramatically with increasing pulse position offset. Hence, VQ of the transform coefficients directly, or a decomposition in the transform domain operating across several pitch cycles, requires consistent placement of the pitch pulses to be effective.

4.4. Formation of the Pitch Track

A block diagram showing the apparatus used to create the pitch track to time-warp a signal to have a constant pitch is depicted in Figure 4.6. From this diagram, three main sections can be identified:

i) Detection of pitch pulse peaks,

ii) Pulsed/Unpulsed classification, and

iii) Pitch track creation.

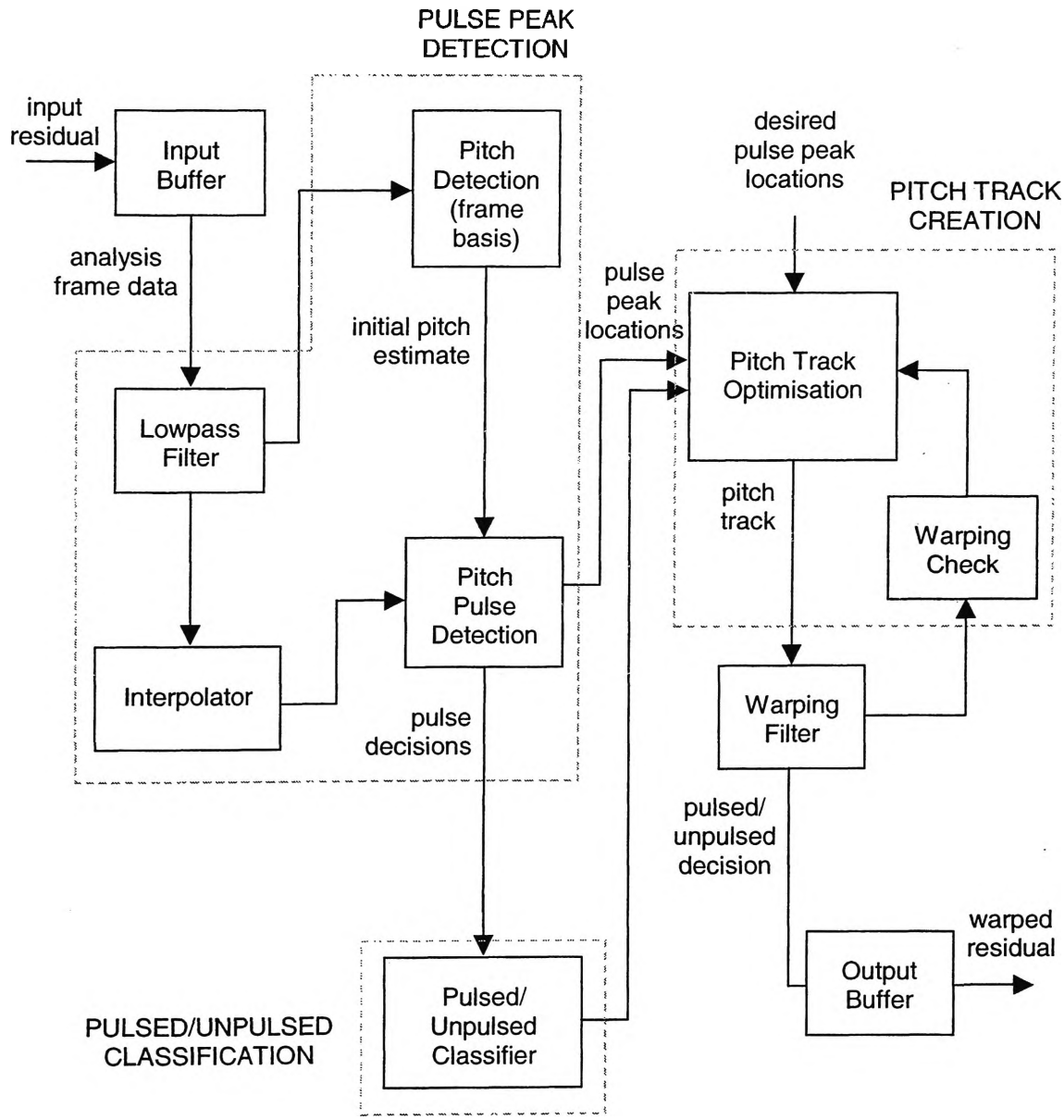


Figure 4.6. Creating the pitch track to ensure pitch pulse peaks of the input residual are warped to the central sample of each warped pitch period

These sections are described in detail below. For the optimisation of the pitch track, two techniques are discussed. The first uses a frame-by-frame basis and applies linear interpolation, while the second operates on a period-by-period basis.

4.4.1. Detection of Pitch Pulses

Autocorrelation Function

The short-time autocorrelation function, *R(d)*, of a length *N* segment of the signal $x(n)$ for a time shift of *d* samples is defined as:

$$R(d) = \sum_{n=0}^{N-1} x(n)x(n-d). \tag{4.4}$$

For a periodic, or quasi-periodic signal, *R(d)* exhibits peaks at time shifts which correspond to multiples of the pitch period, and hence, is commonly used to estimate the pitch of speech as a function of time [Rabi78][Hess83].

Composite Autocorrelation Function

To accurately locate the pitch pulse peaks within a frame, and hence, determine the pitch period value, the residual signal is first lowpass filtered. A pitch detection algorithm, an extension of the technique described in [Haag95], is then applied. The pitch detection method of WMWI, is a two stage process. The first stage determines an initial pitch estimate for the entire frame, then this estimate is refined to a pitch period basis and pitch pulses are located. The pitch estimate for the frame is calculated from the autocorrelations of successive *K* segments, combined to form a composite function. For the case where *K=5*, the composite autocorrelation function, $R_{Composite}$, for each candidate pitch value, *d*, can be expressed as:

$$R_{Composite}(d) = R_3(d) + \sum_{\substack{k=1 \\ k\neq 3}}^{5} a_k \max_{-l(d)\le i\le l(d)} \{w(i)\cdot R_k(d-i)\}, \tag{4.5}$$

where, for segment k, $R_k(d)$ is the autocorrelation function, a_k is a segment weighting factor determined by the pulsed/unpulsed decision of the previous frame, $w(i)$ is a lag-deviation weighting sequence, and $2*l(d)+1$ is the lag-deviation weighting sequence length. Each constituent autocorrelation function is weighted by $w(i)$, centred at the candidate pitch period position, to control the extent that correlation peaks occurring away from the candidate pitch period value contribute to the overall measure. This results in the peaks of R_3 being reinforced by nearby peaks of R_1, R_2, R_4, and R_5, while also preventing unreliable distant peak correlations from skewing the pitch estimate. If some segments are known to be unpulsed, due to decisions made in the previous frame, or have very low signal energy, the correlations for these segments do not contribute to the composite function. The value of d which produces the maximum composite autocorrelation becomes the initial pitch estimate, τ_{init}. This may in fact be a multiple of the pitch period, and hence τ_{init} may be adjusted to a value close to $\frac{\tau_{init}}{m}$, $m \in \{2, 3, 4, 5, 6\}$, if $R_{Composite}$ displays a significant peak at that value.

The constituent autocorrelation functions are then recalculated on an interpolated, filtered residual, for a small set of pitch period values surrounding τ_{init}, using segments of length equal to that value. This is shown by the equation:

$$R_k(d) = \sum_{n=0}^{\tau_{init}-1} x(n)x(n-d), \qquad \tau_{init} - \alpha < d < \tau_{init} + \alpha, \quad k = 1,2,...,K \tag{4.6}$$

where α is the allowable pitch deviation, and $(x(0)\text{-}0.5\tau_{init})$ is the position of the last detected pulse peak. The values $\alpha = 0.2\tau_{init}$ and K=3 are used.

The combination of these K functions, as defined by Equation (4.5), produces a refined pitch estimate for the current period. If the maximum of the refined $R_{Composite}$ exceeds an adaptive threshold, it is proposed that the period contains a pulse. The threshold level is based on the predicted pulse likelihood, determined by the component correlation functions. The pulse peak location is then detected from the interpolated, non-filtered residual signal, providing fractional sample resolution.

Checks are also performed and adjustments made for pitch doubling, pitch halving, and the effects of vocal fry, whereby the pitch period lengthens towards the end of voiced sections of speech. For the case of pitch doubling/halving, the definition of a pitch period as the distance between two adjacent pulse peaks is adhered to. This produces waveforms in the warped domain with a common structure, (single, central pulse) which are suitable for signal decomposition and quantisation.

4.4.2. Pulsed/Unpulsed Classification

The current frame may be classified as pulsed or unpulsed based on the outcome of the refined $R_{Composite}$'s of each pitch period within the frame. The classification is not so much a hard binary decision, but rather one that describes the transition. In other words, if the current frame is classified as pulsed and the previous frame is also

pulsed, it implies that there are continuous pitch pulses throughout the entire current frame. However, if the previous frame was unpulsed, it indicates that a train of pitch pulses begins partway through the current frame. Similarly, an unpulsed classification for the current frame following a pulsed frame implies that the train of pitch pulses diminishes partway through the current frame. The purpose of the pulsed/unpulsed classification is to formulate the structure of the pitch track; it is not used directly for any other purpose. Hence, while the classification has some similarity to a voicing decision, the voiced/unvoiced terminology is not used as this has a particular meaning within phonology and also bears connotations with the method of forming the LP filter excitation. The evolutionary domain decomposition makes a strict voicing decision unnecessary. To eliminate any pitch detection errors due to rapid variation in pitch, significant envelope amplitude changes or erroneous high energy segments of residual, a final series of criteria must be satisfied for the frame to be classified as pulsed. In simple terms, the main criterion for a pulsed frame is that consecutive pitch periods must contain pitch pulses.

4.4.3. Pitch Track Optimisation (Frame Basis)

Optimisation of the pitch track on a frame-by-frame basis as proposed in [Chon99b], whereby the pitch is estimated once per frame and linearly interpolated between updates, involves a two-stage warping process. The second stage refines the pitch estimate of the future frame and re-warps the residual to minimise a measure of deviation between the warped pitch pulse locations and an adaptive set of desired

warped pulse locations. The set of desired warped pulse locations is $\{\alpha, \alpha+L, \alpha+2L,\ldots, \alpha+mL\}$, where α is the warped position of the first pulse of the frame. The method determines the best alignment for all pitch pulses, even after unvoiced or silence periods.

Since the pitch track is linearly interpolated over a frame of length, F, the pitch between two updates (assuming no pitch doubling or halving) can be written as

$$\tau_i = \mathrm{B}i + \mathrm{c}, \qquad i = 0,1,\ldots,F \tag{4.7}$$

where,

$$\mathrm{B} = \frac{\tau_F - \tau_0}{F}, \qquad \mathrm{c} = \tau_0. \tag{4.8}$$

Using this pitch track, the total number of warped samples generated from the first A input samples of the frame is

$$N_{W,A} = \sum_{i=0}^{A-1} \frac{L}{\tau_i} \tag{4.9}$$

where L is the constant pitch period length.

The summation in (4.9) can be well approximated by the definite integral,

$$N_{W,A} \approx \frac{L}{B}\left(\ln\left|A + \frac{c}{B}\right| - \ln\left|\frac{c}{B}\right|\right). \tag{4.10}$$

Likewise, the number of unwarped samples, A, required to generate a given number of warped samples, $N_{W,A}$, can be approximated by the expression,

$$A \approx \exp\left(\ln\left|\frac{c}{B}\right| + N_{W,A}\frac{B}{L}\right) - \frac{c}{B}. \tag{4.11}$$

Standard linear interpolation between pitch updates is performed without regard to the position of the pitch pulses within the frame. Since the main objective in WMWI

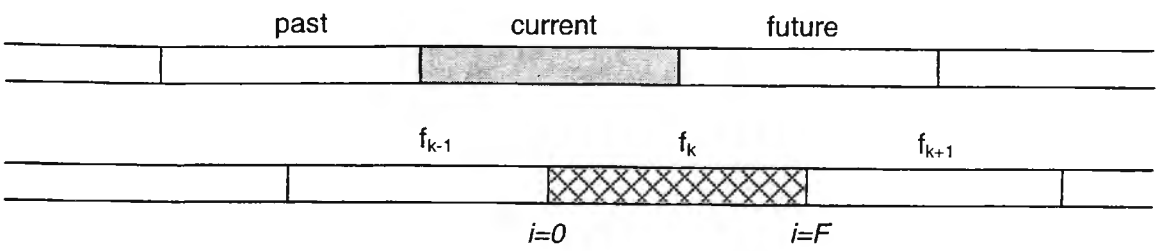
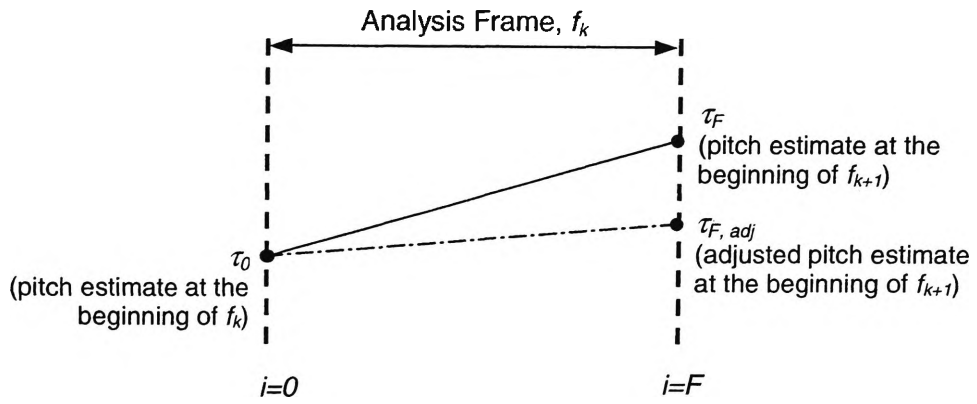


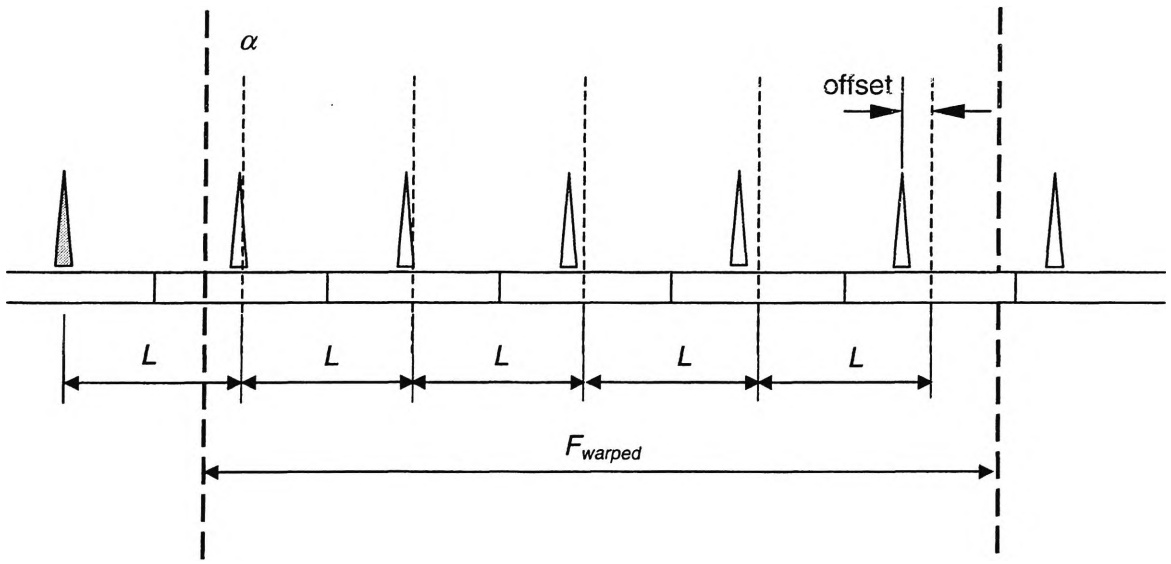
Figure 4.7. Frames used in the frame-based pitch track optimisation method

analysis is to align these pulses, the pitch estimate of the future frame must be modified to enforce correct positioning of the pulses. To determine the pitch track adjustment required, we analyse F samples which lie half a frame ahead of the current frame. This will be referred to as the analysis frame, f_k, as depicted in Figure 4.7. The pitch updates occur at $i=0$ and $i=F$ of the frame f_k, and are denoted τ_0 and τ_F respectively. The central sample of each period, to which pulse peaks are to be warped, is denoted P.

In most cases, the pitch period estimate at the end of f_k, τ_F, is adapted and the new pitch track is derived. The optimisation procedure uses the observation that the accuracy of the linear interpolation approximation degrades as f_k progresses. In other words, if τ_F is incorrect, the distance between the warped pulse locations and the desired warped locations will be larger for pulses towards the end of the frame than for those near the beginning due to accumulating errors, with the last pulse of the frame being offset by the greatest amount from its desired warped location. This is shown in Figure 4.8. The pitch correction procedure required to pull pulses back into alignment, is briefly outlined here.



(a) The initial (solid line) and modified (dashed line) interpolated pitch tracks



(b) The positions of the warped pulses using the non-optimal pitch track. The dotted lines show the desired warped pulse locations.

Figure 4.8. Warping with an interpolated pitch track

1. An initial warping of the residual of f_k is performed using an interpolated pitch track between the current pitch estimate, τ_0, and the future pitch estimate, τ_F, to normalise the pitch period to L samples.
2. If pulses are present within the frame, a new value for τ_F is chosen (this may be a fractional value), such that the last pulse of f_k will be warped to position P. This requires an iterative procedure to solve Equation (4.10) for B (and hence, τ_F). The value τ_F is decremented if the warped pulse location falls short of the desired pulse location, and incremented if the warped pulse location exceeds it. The initial step-size used for the pitch value was one sample, which was then reduced to half and quarter sample resolution. Correct positioning of the last pulse in the analysis frame results in the optimal alignment of all pulses within the frame as a group, under linearly interpolated pitch track conditions. This assumes the pitch varies smoothly. Hence, the positions of the intermediate pitch pulses do not need to be directly considered.
3. If no pulses are present, f_k is declared "unpulsed" and τ_F is set to a nominal predetermined value.
4. If f_{k+1} contains pulses, and f_k is unpulsed, then the pitch track of the last half of f_k is modified to ensure that pulses in f_{k+1} occur at position P of each period. The modification maintains the linear pitch contour, but requires a step-change in the pitch track during the unpulsed frame to ensure exactly (k+0.5), where k is an integer, pitch periods are formed before the first pulse peak (or optimum position for this peak governed by the exact positioning of the last pulse peak). The size of step-change required is calculated using Equations (4.7) to (4.11).

If the pitch track is correct, pulses in the warped domain are spaced approximately L samples apart, and at position P of each period. The method, however, does suffer limitations due to updating the pitch only once per frame during steady state regions and twice per frame during transitional regions. This results in good alignment of the pitch pulses, but not perfect alignment. To increase the warping/alignment accuracy, the pitch is required to be updated more frequently, with perfect alignment achievable if the update rate is once per pitch period.

4.4.4. Pitch Track Optimisation (Period Basis)

Formation of the pitch track on a period-by-period basis enables the exact alignment of pitch pulses in the warped residual. Care must still be taken, though, to ensure that pitch transitions are smooth, and pulses are positioned into alignment after unvoiced sections. Given the pitch pulse locations detected by the methods of Section 4.4.1, the pitch track is then formed. We define the pitch track for a set of four possible frame types:

A. Continuously Pulsed,

B. Continuously Unpulsed,

C. Unpulsed-to-Pulsed, and

D. Pulsed-to-Unpulsed.

It is reiterated, that the primary objective in forming the pitch track within the WMWI framework, is to maximally align pitch pulse peaks to enable good signal analysis and decomposition. This is opposed to other pitch detection algorithms

whose main focus may be to accurately track the changing rate of vibration of the vocal folds.

A. Continuously Pulsed Frame

During a Continuously Pulsed section, a simple, yet effective, technique is to simply allow the pitch to remain constant for the duration of the pitch period. This requires a significantly reduced number of computations in comparison to the iterative procedure used for the linear interpolated pitch track. The pitch period value, $\tau(n)$, associated with samples between two pulse peaks is, in simple terms, equal to the number of unwarped samples between those peaks.

$$\tau(n) = x_{i+1} - x_i, \qquad \begin{matrix} 0 \le i < N_p - 1, \\ x_i \le n < x_{i+1} \end{matrix} \tag{4.12}$$

where x_i is the position (in samples, at fractional resolution) within the unwarped frame of the i^{th} pulse peak, as depicted in Figure 4.9, and N_p is the number of pulses in the frame, including the first pulse of the following frame. The equally-spaced warped domain pitch pulses centred within each period are depicted in Figure 4.10.

This method creates a step-wise pitch track, which is much easier than an interpolated pitch track to control, to guarantee there are exactly L warped samples between pulse peaks. Note, $L = Pf_s$, where P is the pitch period in seconds, as defined in Equation 3 and f_s is the sampling frequency.

Warping Refinement

A refinement of $\tau(n)$ is required to ensure the CWs are perfectly aligned. Consider the effect on an evolutionary domain decomposition of time domain waveforms

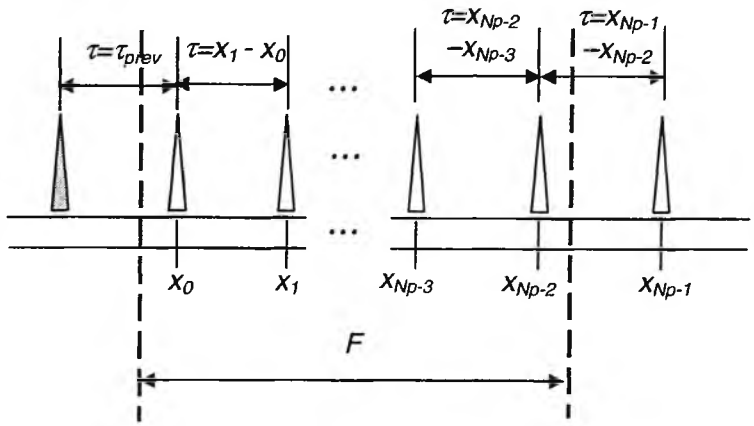


Figure 4.9. Diagram of pitch periods within a "Continuously Pulsed" frame in the unwarped time domain.

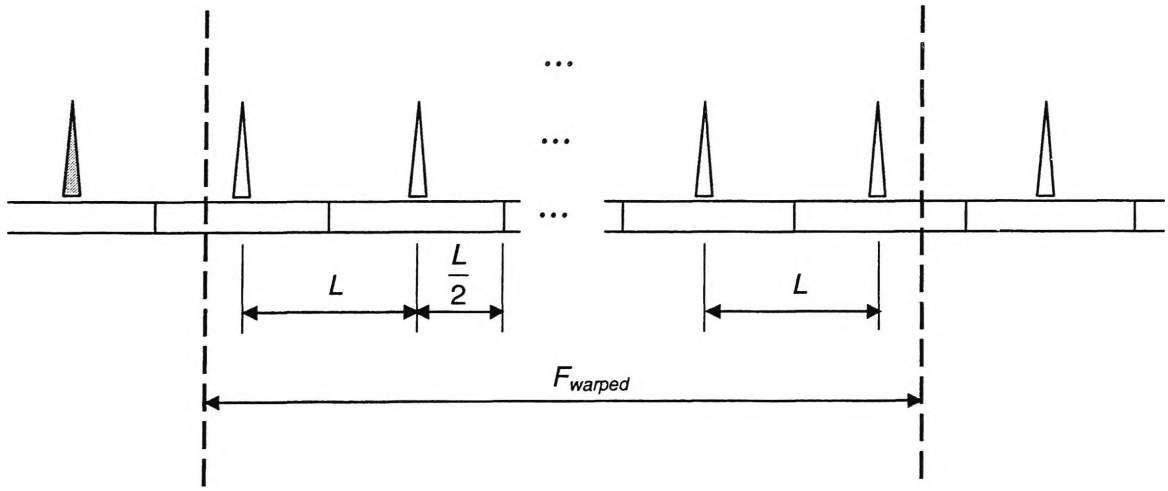


Figure 4.10. Diagram of pitch periods within a "Continuously Pulsed" frame in the warped time domain.

which contain small alignment errors. When values exactly L samples apart are filtered, non-aligned pulses will be in isolation, and hence, decomposed into the rapidly evolving waveform. If pulse peaks are closely, but not exactly aligned, a pulse will still result in the component of rapid evolution, but it will be of lower amplitude. In either case, quantisation of the rapidly evolving component is made very difficult. Due to the nature of the warping filter, it is not possible to determine exactly which of the multiple warped samples generated by a single peak input sample will contain the maximum value. Hence, during pulsed frames, a warping refinement is performed on a period-by-period basis, adjusting the pitch period value by a fractional amount, to correct any minor alignment errors and ensure the distance between pitch pulse peaks is precisely L samples.

B. Continuously Unpulsed Frame

During Continuously Unpulsed frames, the pitch takes on a nominal value, denoted τ_{unv}. Even though the pitch remains constant for the entire frame, the positions of the "pitch period" boundaries are still of interest, as these are required to assist the regaining of alignment as soon as pitch pulses regenerate. This is due to the methods of WMWI, in which alignment of the pulses takes past history into account and cannot simply be achieved through circular rotation of the CWs.

C. Unpulsed-to-Pulsed Frame

For Unpulsed-to-Pulsed frame transitions, the key requirement is to create the pitch track to ensure the upcoming train of pitch pulses following a variable duration unvoiced segment are aligned with those pitch pulses which preceded it. Hence,

the number of whole periods, N_u, and the pitch, τ_u, of these periods comprising the unpulsed section of the frame preceding the period with the first pulse, must be chosen such that the first pulse peak, located at x_i, is warped to the correct position.

To minimise the pitch variation, we solve

$$\{N_u, \tau_u\} = \underset{\substack{N_u' \\ N_u'=1,2,3,...}}{\arg\min} \left| (x_2 - x_1) - M\tau_u \right| \tag{4.13}$$

where,

$$M\tau_u = \frac{x_1 - \frac{x_2 - x_1}{2} - y}{N_u'}, \qquad \begin{matrix} N_u' = 1,2,3,.... \\ \tau_{u\min} < \tau_u < \tau_{u\max} \end{matrix} \tag{4.14}$$

where x_i is the position of the i^{th} pulse peak, y is the boundary of the last period of the previous (unpulsed) frame, and M is the interpolation constant.

The placement of period boundaries for an unpulsed-to-pulsed frame of (unwarped) residual is shown in Figure 4.12. In the case depicted, the number of pulses, N_p, is 4 (one pulse lies in the future frame which is not shown) and N_u=3. The pitch of these unpulsed periods is selected using Equation (4.13) to allow the peak at x_1 to be centred within the fourth period. If x_1 is very close to the beginning of the frame, Equation 2 may be indeterminate due to the constraints on τ_u. If this occurs, y is shifted back to the previous period boundary, and τ_u is recalculated. The pitch of the "pulsed" part of the frame is determined using the methods of Continuously Pulsed frames. The pitch for each sample in the frame can then be expressed as follows:

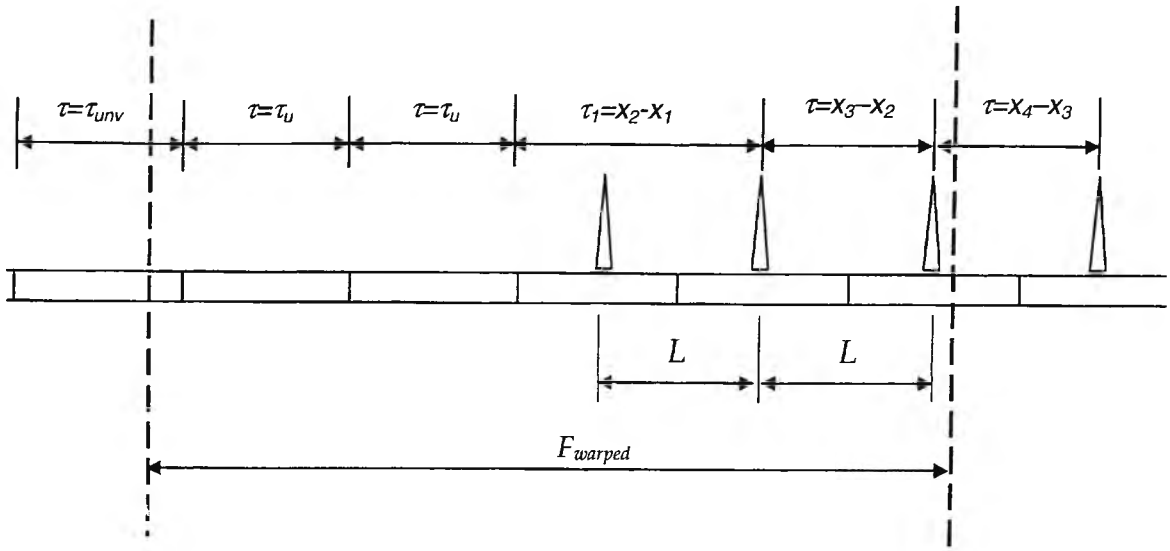


Figure 4.11. Diagram of pitch periods within a "Unpulsed-to-Pulsed" transition frame in the warped time domain, in the case where the number of periods preceding the period with the first pulse is 2.

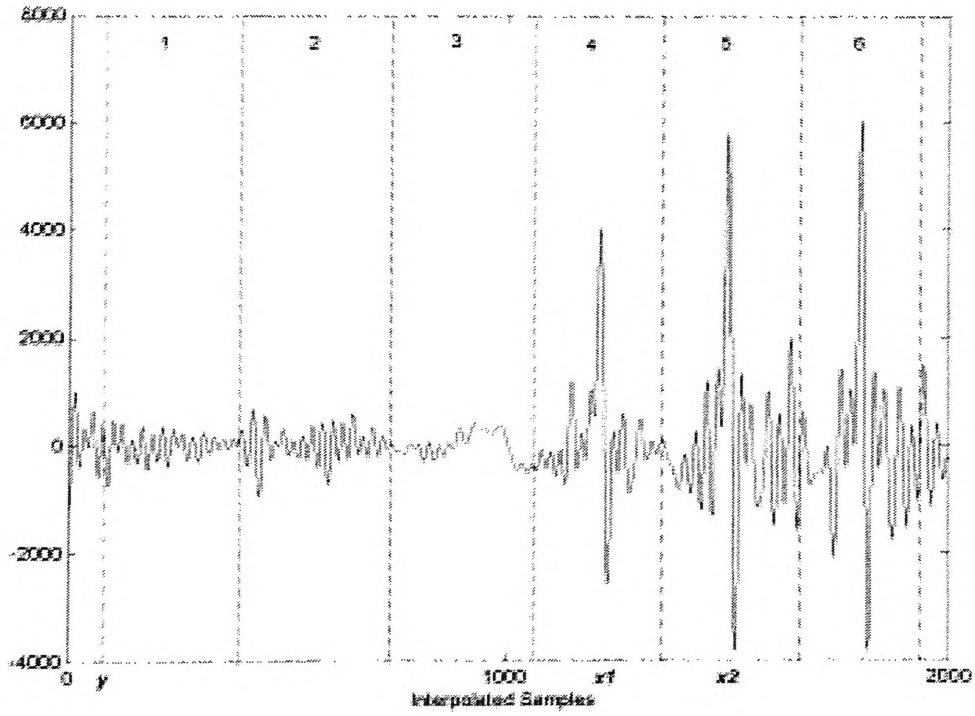


Figure 4.12. The placement of pitch period boundaries in a frame with an unpulsed-to-pulsed transition. The pitch of periods 1-3 is selected to allow the peaks of period 4 to be centred within the pitch period.

$$\tau(n) = \begin{cases} \tau_{unv}, & 0 \le n < y \\ \tau_u, & y \le n < x_1 - \dfrac{\tau_1}{2} \\ \tau_1 = x_2 - x_1, & x_1 - \dfrac{\tau_1}{2} \le n < x_1 \\ \tau_i = x_{i+1} - x_i, & x_i \le n < x_{i+1} \end{cases}, \quad 0 \le i < N_p \tag{4.15}$$

D. Pulsed-to-Unpulsed Frame

During Pulsed-to-Unpulsed frames, the pitch of the pulsed part of the frame is calculated using the adjusted and refined results of Equation (4.12). The remainder of the frame takes on the predetermined "unvoiced" pitch value. Since unpulsed speech does not contain quasi-periodic features and is more random in nature, less care is required in the transition to unpulsed speech than for the transition to pulsed speech.

The pitch track calculation places a great deal of importance on the positions of the pitch pulse peaks, and the locations of the pitch period boundaries. This is necessary to achieve effective decomposition and quantisation without compromising the waveform coding objective.

Summary

Formation of the pitch track for signal analysis is best performed on a pitch period basis, rather than the frame basis previously described. It should be noted, however, that while period-by-period pitch track analysis is preferable to obtain well-aligned CWs for effective quantisation, the exact analysis pitch track does not need to be transmitted to the decoder, and hence may not require a large increase in the bit rate. If the required bit rate is low, a single pitch value, such as the refined

pitch estimate of the frame-basis technique, or a pitch value optimised in favour of the adopted reconstruction technique, could be transmitted and linearly interpolated to reconstruct the waveform. This is explained in more detail in Section 5. It is acknowledged that this method of pitch track formation is more complex than others commonly used, but the accuracy and benefits obtained by the warped signal representation enable waveform coding, eliminate the need for rotational alignment, and also achieve computational and transmission rate savings in subsequent procedures.

4.5. Transforms

Normalisation of the pitch allows fixed length transforms to be performed on each pitch cycle at a constant rate. The transforms may offer benefits such as lower entropy (tightly clustered coefficients) or slower evolution of coefficients (smaller bandwidth) which lead to higher coding efficiency. Efficiency is measured in terms of the transform coding gain, defined as

$$G_{TC} = \frac{\sigma_{q,PCM}^2}{\sigma_{q,TC}^2}, \tag{4.16}$$

where $\sigma^2_{q,PCM}$ is the quantiser noise variance using PCM, and $\sigma^2_{q,TC}$ is the quantiser noise variance for the transform with optimal scalar quantisation and bit allocation. The transforms discussed can be interpreted as perfect reconstruction filter banks, and by choosing the window length to be a power of two, they may be implemented with fast algorithms [Malv92].

4.5.1. Block Transforms

With block transforms, the input signal is divided into "blocks" of equal length which do not overlap, and each block is transformed. In the WMWI framework, the block length is equal to the constant pitch period, L, and hence, transforms may be applied pitch-synchronously to the warped residual. Transforms commonly used are the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT) [Malv92], and Gabor Transform (GT) [Qian93]. These transforms may be implemented as filter banks with N channels, filters of length N and down-sampling by N. However, block transforms suffer from energy discontinuities at the block boundaries, known as "blocking effects", which produce audible artefacts, especially if rectangular windowing is used. Use of windowed exponentials in the GT produces the desired smoothing, however, this is at the expense of critical-sampling. The GT, with an oversampling factor of 2, is defined as

$$b_k(m) = \sum_{n=0}^{L-1} s\left(n + \frac{L}{2}\right) w(n) W^{-kn},$$

$$s\left(n + \frac{L}{2}\right) = \frac{1}{4} \sum_{k=0}^{L-1} \sum_{n=0}^{\frac{L}{2}-1} b_k(m) g(n) W^{kn} + b_k(m-1) g\left(n + \frac{L}{2}\right) W^{kn}, \tag{4.17}$$

where $w(n)$ is the hamming analysis window, $g(n)$ is the anti-aliasing synthesis window, and W^{kn} denotes the exponential modulation

$$W^{kn} = \exp\left(\frac{j2\pi kn}{L}\right). \tag{4.18}$$

4.5.2. Lapped Transforms

Lapped transforms have a number of advantages over block transforms, including higher coding gains, no blocking effects and better frequency resolution. A popular class of lapped transforms are the cosine modulated transforms [Malv92][Vaid93]. In WMWI analysis, these may be implemented as L-channel filter banks, where all L analysis filters are derived from a single prototype filter by cosine modulation. In the case where the analysis filter is restricted to have a filter length of $2L$, perfect reconstruction may be achieved and the transform can be viewed as a generalisation of the Lapped Orthogonal Transform (LOT) [Malv92]. The lapped transforms therefore provide a pitch-synchronous subband representation of the speech. The input is decomposed by the cosine modulated filters, then critically down-sampled. A diagram of the analysis/synthesis system is shown in Figure 4.13 and the approximate magnitude response of the modulated filter bank is in Figure 4.14. Polyphase implementations can be used to reduce the computational load.

The modulated filter bank is given by

$$f_k(n) = h(n)\sqrt{\frac{2}{L}}\cos\left(\left(n+\frac{L+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{L}\right), \tag{4.19}$$

where k is the overlap factor, L is the transform length, and the basis function, $h(n)$, is

$$h(n) = -\sin\left(\left(n+\frac{1}{2}\right)\frac{\pi}{2L}\right), \quad n = 0,1,2,\ldots,2L-1 \tag{4.20}$$

for the Modulated Lapped Transform (MLT) ($k = 1$), and

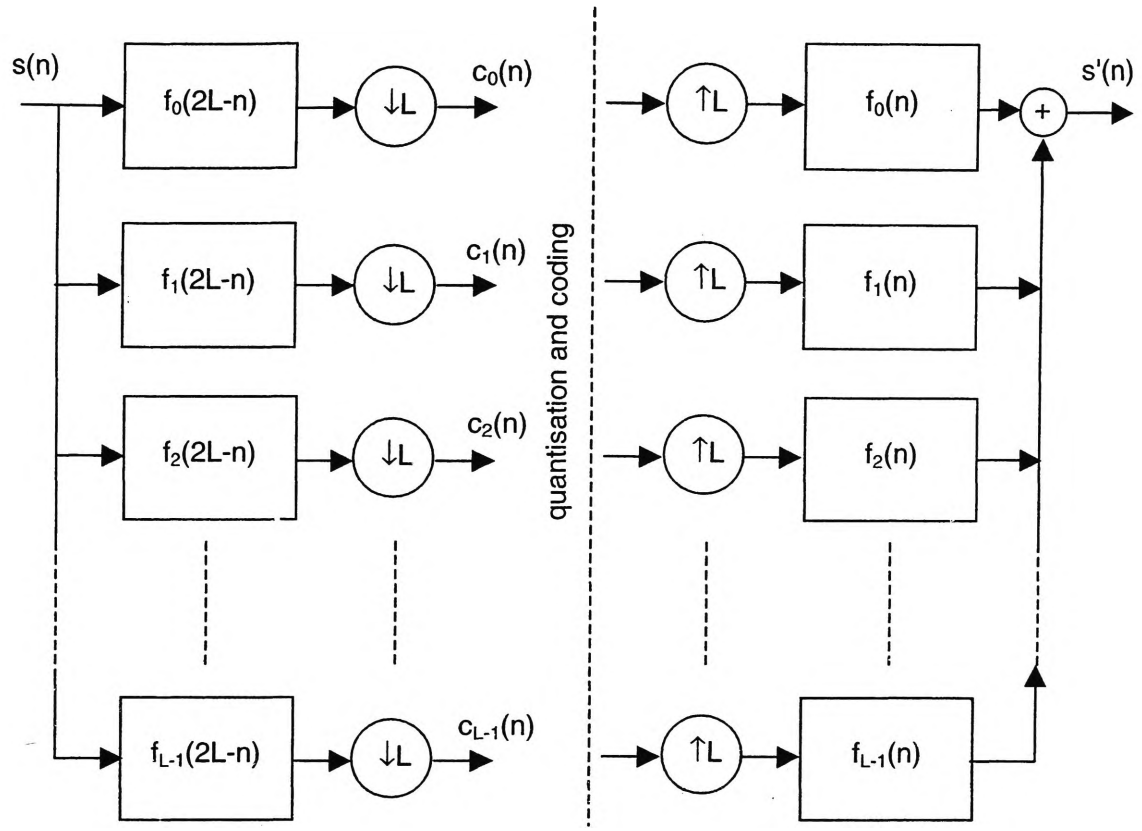


Figure 4.13. The analysis/synthesis filter bank interpretation of the MLT.

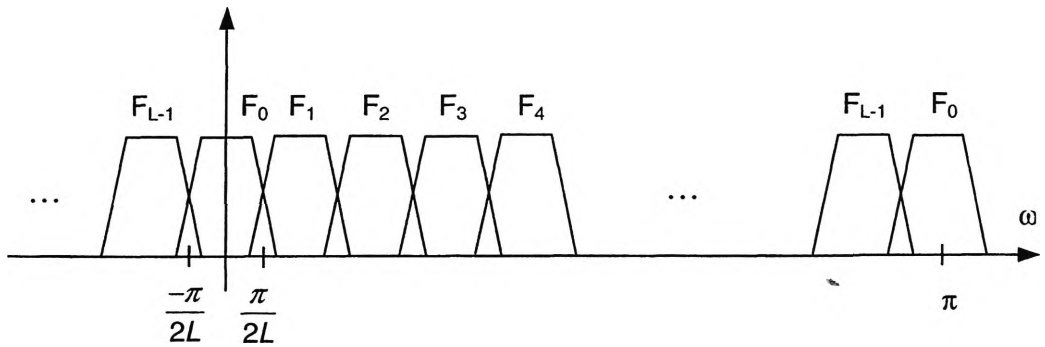


Figure 4.14. Magnitude response of the filter bank of the MLT.

$$h(n) = -\frac{1}{2\sqrt{2}} + \frac{1}{2}\cos\left(\left(n+\frac{1}{2}\right)\frac{\pi}{2L}\right), \quad n = 0,1,2,...,4L-1 \tag{4.21}$$

for the Extended Lapped Transform (ELT) (k = 2).

The MLT is defined by

$$\begin{aligned} c_k(m) &= \sum_{n=0}^{2L-1} s(n+mL) f_k(n), \\ s(n+mL) &= \frac{2}{L} \sum_{k=0}^{L-1} \sum_{n=0}^{L-1} c_k(m) f_k(n) + c_k(m-1) f_k(n+L) \end{aligned} \tag{4.22}$$

Note that critical sampling is implied in the transform definition.

While the MLT is preferred over the LOT due to its slightly higher coding gain and lower computational complexity, the ELT has an even higher coding gain than the MLT, higher stopband attenuation, and also higher SNR [Malv92]. This, however, is at the expense of a longer delay.

4.5.3. Effect Of Warping On The Transforms

Warping (or stretching) of the signal in the time domain corresponds to the shifting of frequency components towards zero (compression). While the transform of each warped pitch period produces L coefficients, the number of *significant* transform coefficients is approximately equal to the pitch period for real-valued transforms. Higher frequency coefficients have values close to zero, and therefore, truncation of the coefficient series at *l=pitch* can be performed with little distortion. Hence, fixed-length vector quantisation techniques can be used for the time-domain waveforms, but in the transform domain, the coefficient series are, in essence, still variable-length, necessitating variable dimension VQ.

The perfect reconstruction filter banks are associated with disadvantages such as increased delay and, in the case of lapped transforms, coefficients which describe the properties of more than one pitch cycle. This makes it difficult to apply auditory masking models to ease the quantisation task. It was perceived that the transforms did not produce significant coding advantages which justified the increased delay and additional computational cost. Greater merit is seen in the simpler decomposition methods of previous WI coders to improve quantisation efficiency. Other transforms may lead to advantageous decompositions, e.g. [Klei00], though this is not currently apparent.

4.6. Decomposition Techniques

In WMWI, the decomposition of the CW surface into perceptually different components may be performed on the critically-sampled CWs with a variable-tap filter whose length depends on the pitch. Alternatively, the surface may be resampled to produce a constant sampling rate, which allows the use of fixed-coefficient filters with constant delay. The resampling from pitch-synchronous rate to a fixed sampling rate is performed with B-spline interpolation [Unse93] of the evolving surface, creating a fixed number of 10 CWs/frame. Several decomposition methods are discussed here:

A. Fixed lowpass filtering,

B. Adaptive/Switched lowpass filtering,

C. Differential decomposition, and

D. The pitch synchronous wavelet transform (PSWT)

The decomposition is performed on time domain waveforms. This choice of domain was motivated by the requirements of the subsequent quantisation techniques, detailed in Section 5.5, whereby the SEW component, in particular, is able to be better quantised in the warped time domain than in the frequency domain.

A. Fixed lowpass filtering

In common WI, lowpass filtering decomposes the evolution of the CW surface, shown by the block diagram in Figure 4.15. The filter output is the slowly evolving waveform (SEW) and the difference between the CW and the SEW is labelled the rapidly evolving waveform (REW).

This decomposition generates a very smooth SEW surface with a 20th order 20Hz, lowpass FIR filter, but has the disadvantage of being slow to react to unvoiced-voiced transitions. The SEW filter tends to smooth out any sudden changes in the signal evolution, gradually building up to large pulses, rather than representing the actual instantaneous increase in the signal energy. This causes the SEW to contain pulses, due to the foresight of the filter, when the corresponding segment of the CW is noisy (Figure 4.18). In addition, at the beginning of a pulsed section, the low

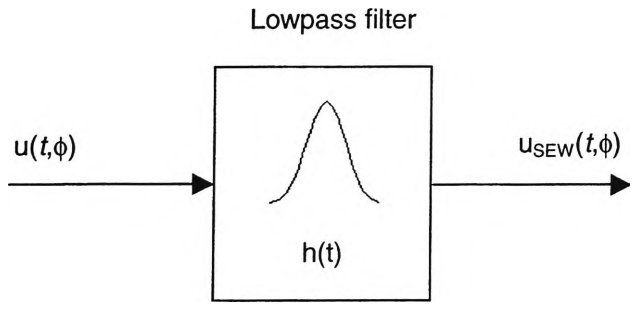


Figure 4.15. Common lowpass filtering decomposition technique

amplitude noise preceding the pulse train decreases the amplitude of correct SEW pulses. The REW then compensates for these errors, resulting in (negative) pulses to null out the false SEW pulses, as well as higher energy than desired at the beginning of periodic sections to reinforce the underestimated SEW pulses. This is unfavourable for REW quantisation.

The SEW gain may be artificially adjusted to prevent the SEW energy exceeding the CW energy, however, to allow effective quantisation, it is preferable for the pulse shape to be eliminated from the REW completely.

B. Adaptive Lowpass Filtering

WMWI has the advantage, due to its pitch pulse detection algorithm, of knowing in advance when a series of voiced pulses starts and finishes. This information can be used to enable a decomposition technique which adapts during transitional frames (unpulsed-to-pulsed or pulsed-to-unpulsed) depending on the start/end locations of the pulse train. To solve the problem of smoothed pulse transitions and false evolution shapes, a possibility is to use a switched decomposition filter (Figure 4.16).

For example, in steady state regions, a longer, higher *frequency* resolution filter is used, but in transition regions, a shorter, higher *time* resolution filter is used. Such techniques are utilised in audio compression to combat exactly the same problem, e.g. [Sinh96]. Based on the time-varying characteristics of the signal, the filter is switched between one which dynamically tracks rapid changes and one which produces a smoothly evolving surface. This enhances the quality of signal

segments containing rapid energy changes. Alternatively, rather than switching the filter, the same SEW filter could be maintained, but the amount that each of the filter taps contributes to the output could be adjusted to prevent future (or past) pulses influencing the current filter output (Figure 4.17). The weighting is expressed in the adaptive function *w(t)*. The result is shown in Figure 4.19 and compared to the common WI lowpass filter output of Figure 4.18. Note that unlike the decomposed SEW of the common filtering technique, the pulse train for the adaptive filtering method does not begin in the SEW until the first pulse is present in the residual

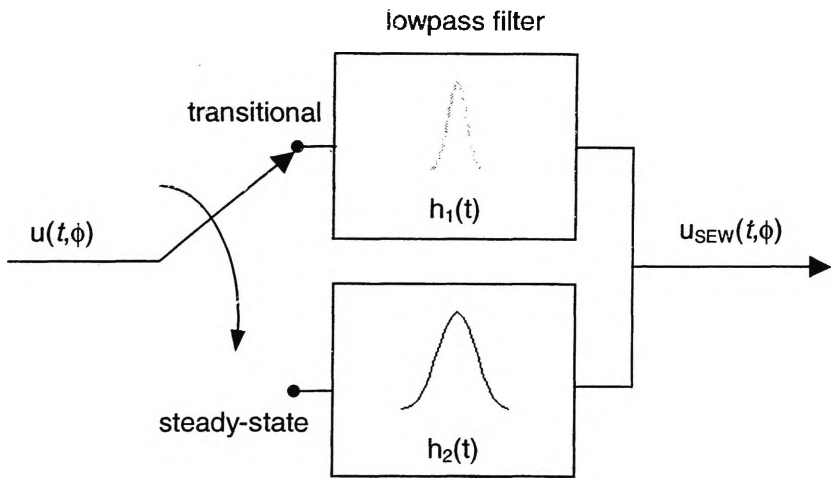


Figure 4.16. Switched decomposition filter

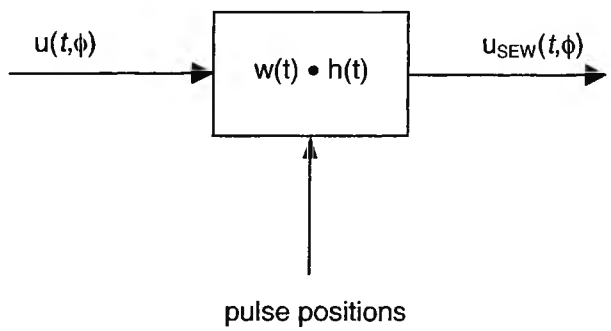


Figure 4.17. Adaptive decomposition filter

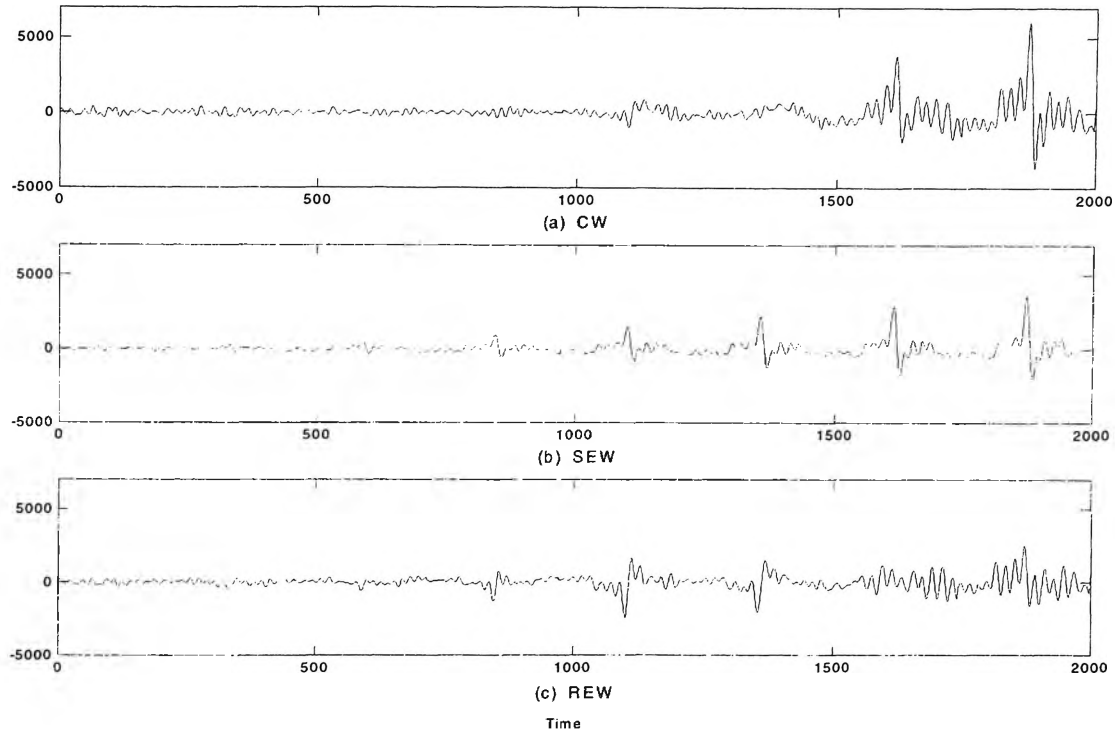


Figure 4.18. Decomposition using the common WI lowpass FIR filter. Note how the REW contains pulses to null out the incorrect SEW pulses.

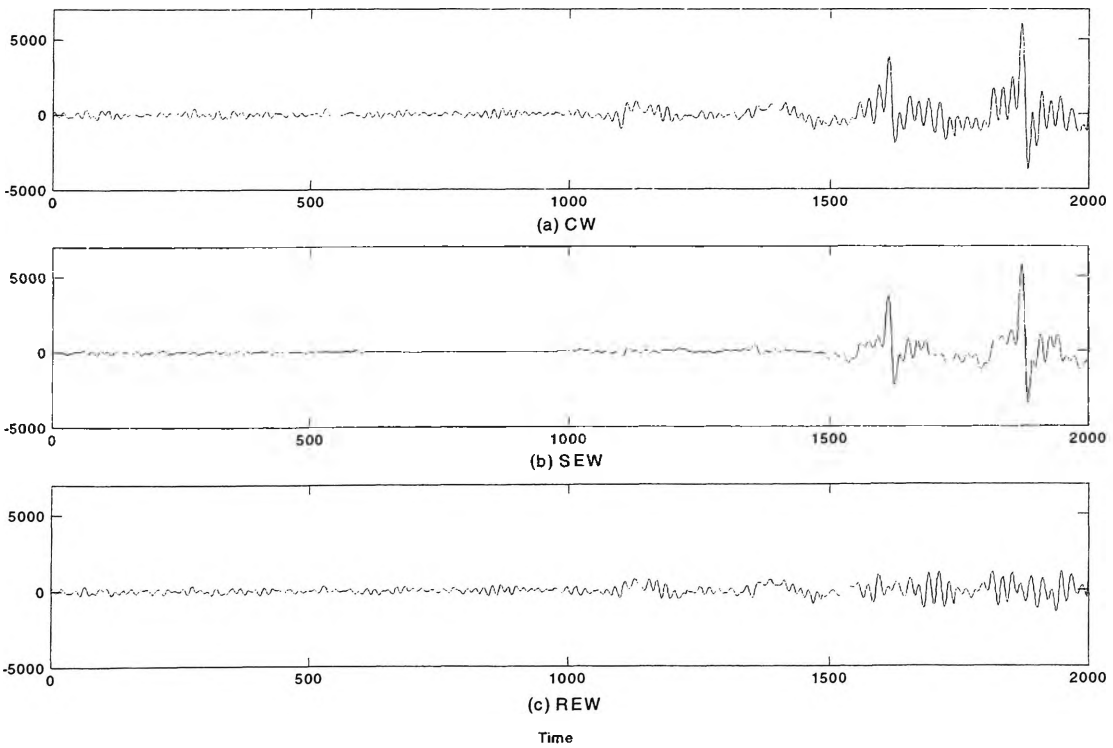


Figure 4.19. CW decomposition using an adaptive lowpass FIR filter. The SEW correctly represents the quasi-periodic component of the CW.

signal. This leads to an improved representation of the periodic and non-periodic components of the residual during transitional regions.

C. Differential Decomposition

An alternative, differential technique was developed to improve the SEW/REW representation during transition regions. This decomposition, shown in Figure 4.20, is performed on critically sampled pitch periods, and requires past, present and future pitch cycles, $x_{n-1}(k)$, $x_n(k)$ and $x_{n+1}(k)$ respectively. In this method, the optimum gains, λ_1 and λ_2, between $x_{n-1}(k)$ and $x_n(k)$, and $x_n(k)$ and $x_{n+1}(k)$ respectively are determined such that the Mean Squared Error(MSE) between the current CW and gain adjusted adjacent CW is minimised.

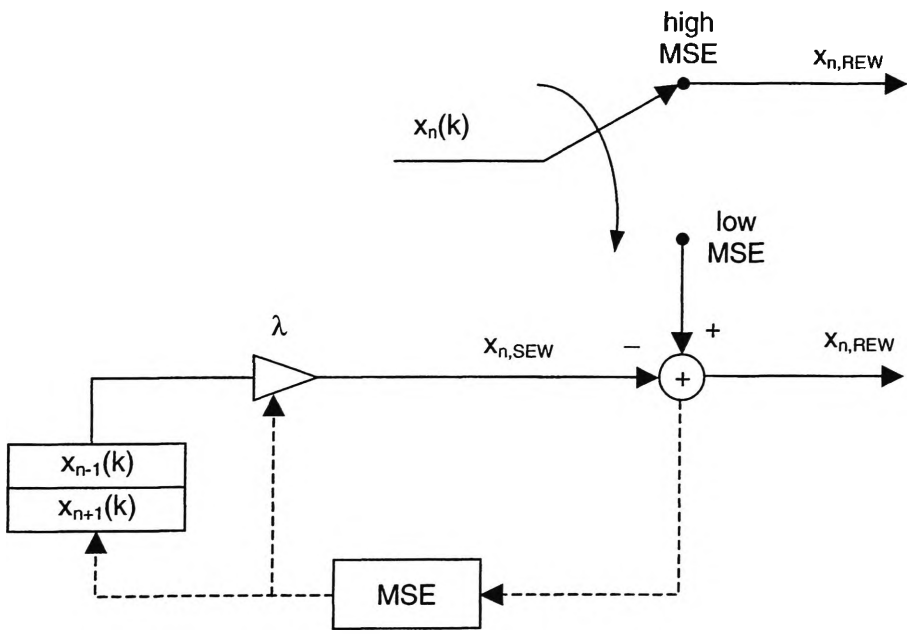


Figure 4.20. Differential decomposition technique

This is calculated as follows:

$$MSE_1 = \sum_k [x_n(k) - \lambda_1 x_{n-1}(k)]^2 \tag{4.23}$$

$$MSE_2 = \sum_k [x_n(k) - \lambda_2 x_{n+1}(k)]^2 \tag{4.24}$$

$$\lambda_1 = \frac{\sum_k x_n(k)x_{n-1}(k)}{\sum_k x_{n-1}^2(k)} \tag{4.25}$$

and likewise,

$$\lambda_2 = \frac{\sum_k x_n(k)x_{n+1}(k)}{\sum_k x_{n+1}^2(k)} \tag{4.26}$$

Substituting (4.25) into (4.23),

$$\begin{aligned} MSE_1 &= \sum_k x_n^2(k) - 2\frac{\sum_k x_n(k)x_{n-1}(k)}{\sum_k x_{n-1}^2(k)} \sum_k x_n(k)x_{n-1}(k) \\ &\quad + \left(\frac{\sum_k x_n(k)x_{n-1}(k)}{\sum_k x_{n-1}^2(k)}\right)^2 \sum_k x_{n-1}^2(k) \\ &= \sum_k x_n^2(k) - \frac{\left(\sum_k x_n(k)x_{n-1}(k)\right)^2}{\sum_k x_{n-1}^2(k)} \end{aligned} \tag{4.27}$$

Hence, to find min(MSE_1, MSE_2), we must solve

$$\max\left(\frac{\left(\sum_k x_n(k)x_{n-1}(k)\right)^2}{\sum_k x_{n-1}^2(k)}, \frac{\left(\sum_k x_n(k)x_{n+1}(k)\right)^2}{\sum_k x_{n+1}^2(k)}\right) \tag{4.28}$$

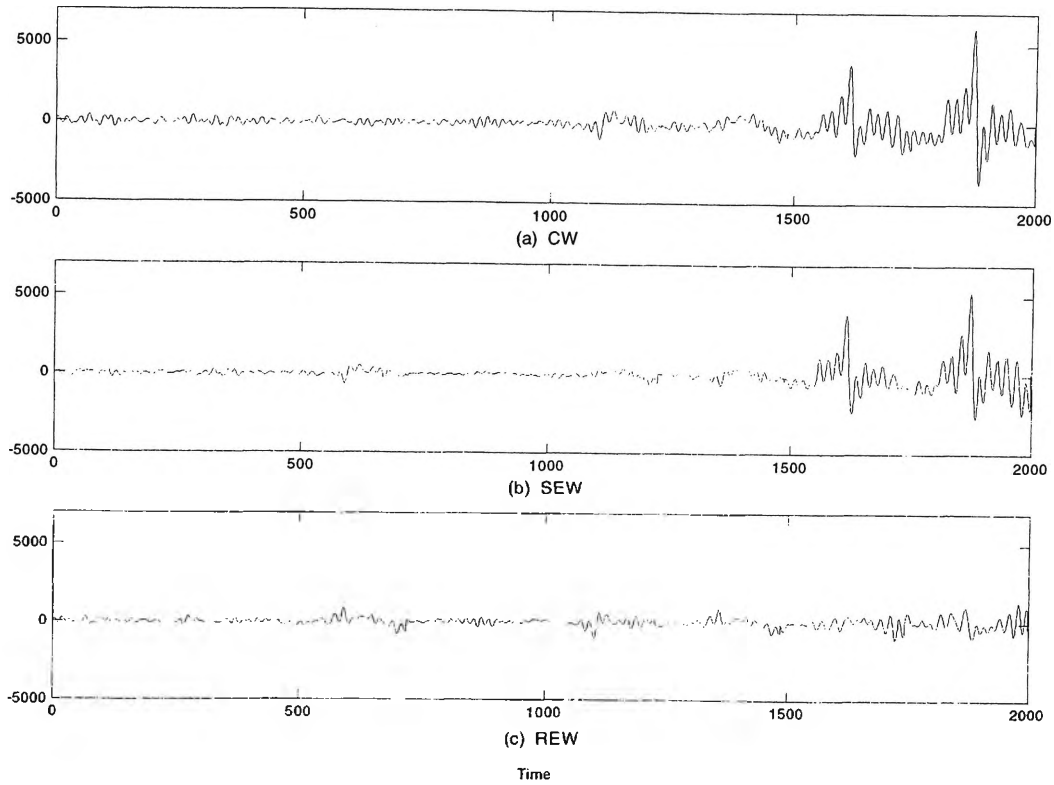


Figure 4.21. CW decomposition using the differential decomposition.

The gain-adjusted past or future pitch period is then subtracted from the current CW, depending on which gives us the minimum MSE. The remainder is the noise-like "REW" component and the subtracted portion is comparable to the "SEW" component.

The decomposition works well for pulsed sections, however, in unpulsed, noisy sections, identified by a high MSE, there is little correlation between the current and adjacent periods. In these cases, the slowly-evolving component is insignificant; the period is labelled as noise and all energy is transferred to the "REW". A section of speech residual decomposed using the differential decomposition is shown in Figure 4.21.

D. Pitch Synchronous Wavelet Transform (PSWT)

The wavelet decomposition based on the PSWT described in Section 3 can also be applied to the constant pitch CW surface. This method provides a multi-resolution decomposition of the CW evolution, producing a slowly evolving component, as well as characterising the rapidly evolving component at several scales. The PSWT is implemented as a tree-structured filter bank of the form depicted in Figure 4.22. The difference between the wavelet decomposition described in the previous chapter and the decomposition proposed here, is that the warping operation, coupled with the accurate pitch detection algorithm, allows the transform to be performed on critically-sampled pitch periods, as intended. Such a method to form the pitch-synchronous representation in real time had not previously been established, requiring the decomposition to be performed on oversampled, rotated pitch cycles.

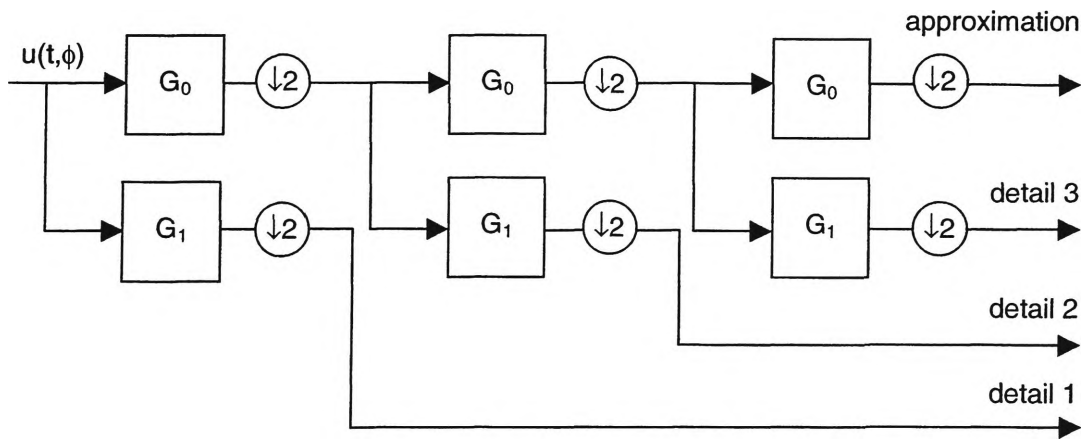


Figure 4.22. Decomposition based on the PSWT, where filter G_0 is lowpass and G_1 is highpass

4.6.1. Delay

The delay incurred by Decompositions A and B is equal to *(N+1)/2* characteristic waveforms, where *N* is the filter length. Since this delay is in the evolution domain, and the CW surface has been resampled to give ten equally-spaced CWs for each frame, a 20th order FIR filter incurs a delay of one frame, or 25ms.

Since Decomposition C requires one future (critically-sampled) pitch period, the delay is variable in the unwarped time domain with a maximum delay incurred of 147 samples (the maximum pitch period). Alternatively, only past pitch periods could be used in the calculation and hence, no delay is incurred. Decomposition D incurs a delay of $(2^N-1)L$, where L is the combined group delay of the analysis/ synthesis pair, and N is the number of decomposition levels used.

The preferred technique is Decomposition B, the adaptive lowpass filtering technique which uses information regarding the time when pulse onsets begin. This method gives an accurate separation of pulse and noise during transitional regions, while producing smooth SEW surfaces during steady state regions. Also, comparison of Figure 4.19 and Figure 4.21 shows that Decomposition C has the tendency to separate too much of the CW random variations into the SEW, making this component more difficult to quantise.

4.7. Summary

The Waveform-Matched Waveform Interpolation (WMWI) technique aims to improve the quality of conventional WI coders at higher bit rates, by incorporating

the waveform coding property. This is achieved by time-warping the linear prediction residual signal to enforce a constant pitch period, and extracting critically-sampled, fixed-length characteristic waveforms to form a two-dimensional surface. This provides for improved signal analysis over common WI coders, without errors due to cyclic rotation or the repetition or omission of segments by selective extraction. A diagram of the WMWI analysis and decomposition architecture described in this chapter is shown in Figure 4.23.

For effective decomposition and quantisation of the signal using WMWI, the CW surface must contain aligned pitch periods. This requires accurate pitch pulse detection, performed using weighted composite correlation functions, and careful derivation of the pitch track. Two pitch track optimisation techniques were

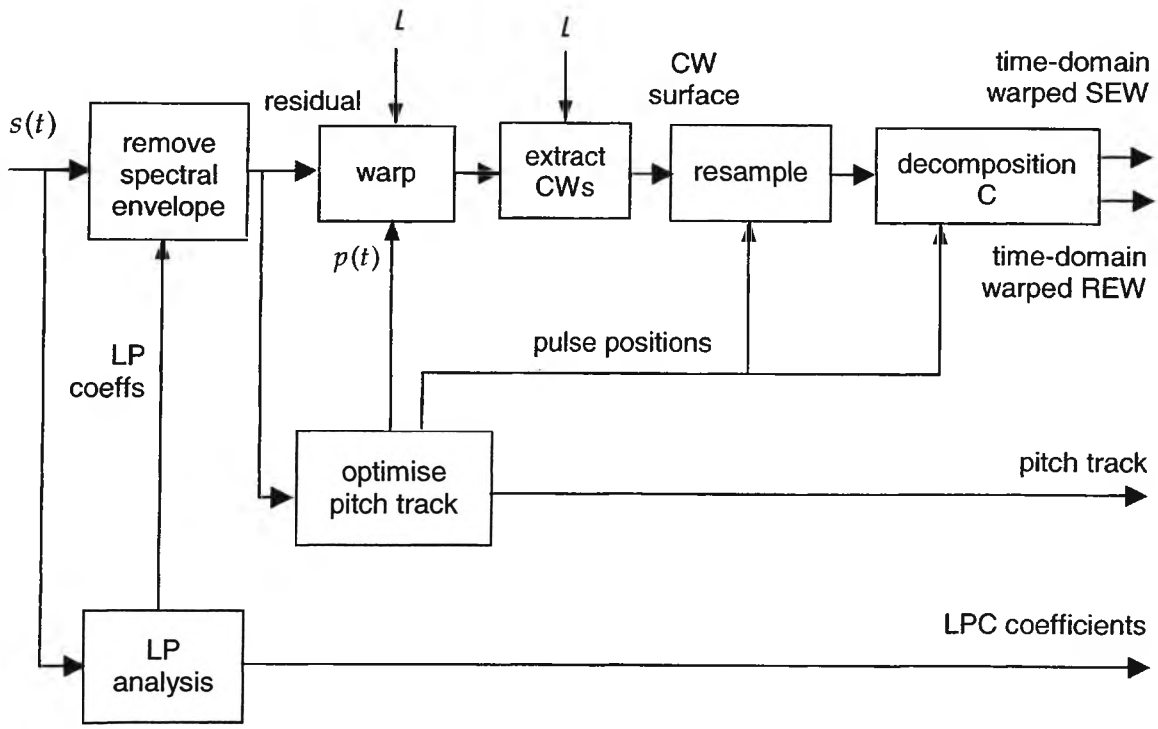


Figure 4.23. WMWI analysis and decomposition

discussed: one frame-based, and the other, pitch period-based. To gain the required alignment accuracy, the pitch track should be created on a pitch period-by-pitch period basis; a method for which was defined for the four possible frame types, based on their pulsed/unpulsed classifications. The technique presented ensures consistent and reliable positioning of the pitch pulses in the warped time domain, and also regains alignment after unvoiced segments.

Once the CW surface has been created, an increase in coding efficiency can be achieved by using a decomposition or transformation to separate the signal into components which are more easily quantised. Several signal transformations and decomposition techniques were discussed. The preferred technique is a modified SEW/REW decomposition which incorporates adaptive lowpass filtering. In this method, the filter taps are weighted during transitional periods, based on the time-locations of pitch pulse onsets, to avoid smoothing effects.

The advantage of WMWI is that no information is destroyed by the analysis techniques. This enables an accurate description of the signal evolution and the ability to obtain near-perfect reconstruction of the speech in the unquantised case.

Chapter 5

Quantisation and Reconstruction of WMWI Parameters

I've come loaded with statistics, for I've noticed
that a man can't prove anything without statistics.

-- Mark Twain

5.1. Introduction

Most low bit rate speech coders are based on lossy speech models, which limit their performance at higher transmission rates. To achieve good scalable performance, it is advantageous for the coder to fulfil the waveform matching property; that is, the output performance converges to perfect quality with increasing bit rate. The WMWI analysis techniques described in the previous chapter are compatible with waveform coding, as they are completely invertible. Hence, the input speech may be precisely recovered at the decoder.

In this chapter, the methods used to quantise and reconstruct the decomposed CW components and the pitch parameter, in order to create an accurate representation of the residual signal, are described (Figure 5.1). In particular, significant emphasis

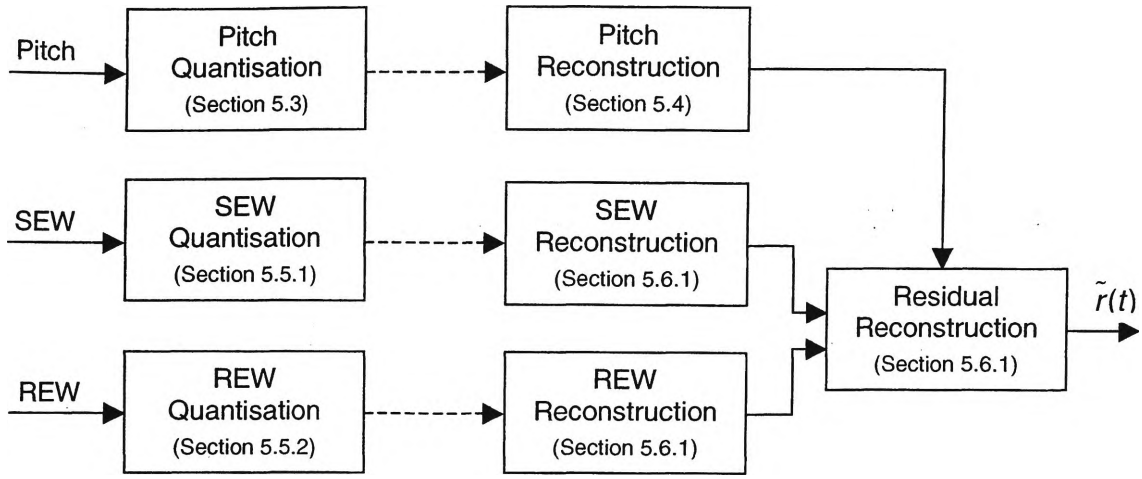


Figure 5.1. Quantisation and reconstruction of WMWI analysis parameters to form the residual signal

is placed on the pitch quantisation and reconstruction scheme, the accuracy of which determines the degree of time-synchrony between the input and reconstructed speech signals. A 4kbit/s implementation of the WMWI coder is described and its subjective quality results are also presented.

5.2. Reconstruction Approaches

Reconstruction of the residual signal of WMWI from the quantised CWs may be performed by two different methods; one which renders an approximate representation of the input speech, and the other which produces an accurate representation. The type of reconstruction performed is governed primarily by the level of similarity between the original and reconstructed pitch tracks, as depicted in Figure 5.2. The advantages and disadvantages of each method are outlined below.

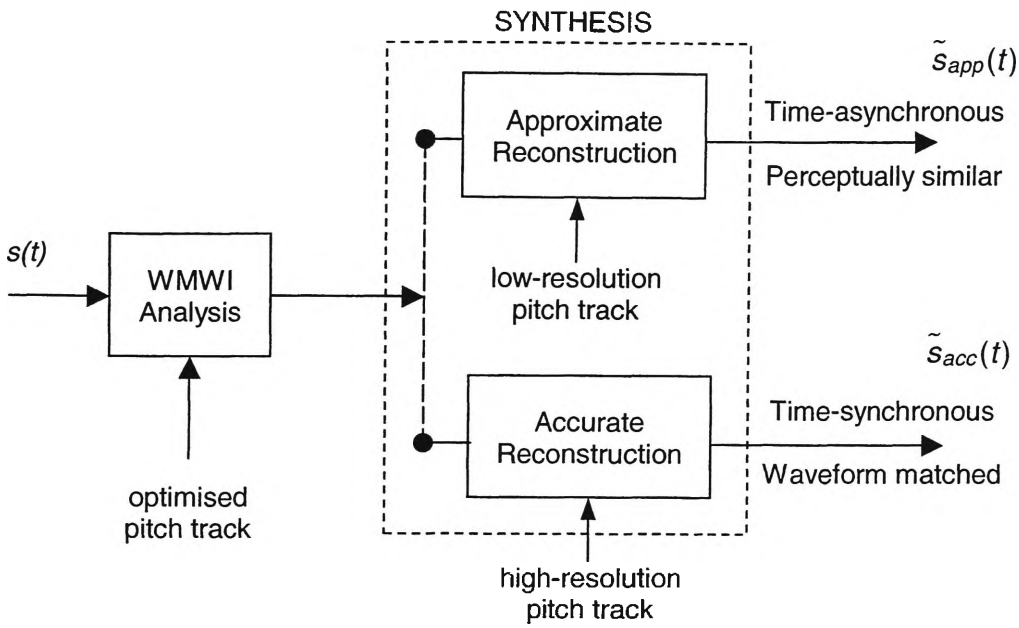


Figure 5.2. The reconstruction choices for WMWI

5.2.1. Approximate Reconstruction

The signal reconstruction method used in standard WI is an approximate reconstruction, whereby the main objective is to produce a signal which sounds perceptually similar to the input signal. It uses a low-resolution pitch track, updated every 20-25ms, and continuously interpolates between the reconstructed CWs without regard to the original positioning of these periods within the frame. This is discussed in more detail in Section 5.4.5. The technique compromises the waveform coding objective as it does not attempt to match the input signal exactly. This makes it unfavourable for the proposed WMWI coder. However, when combined with the WMWI analysis procedures, the accurate signal analysis provides advantages over standard WI, making this type of reconstruction suitable at lower bit rates.

5.2.2. Accurate Reconstruction

An alternative method is the accurate reconstruction technique, which maintains the phase relationships of the CWs by preserving the time-locations of pitch pulses. In this method, the residual is reconstructed by directly inverting the analysis operations, the most significant of which is the unwarping procedure. This technique is the preferred approach for WMWI synthesis as it allows waveform matching. It should be noted that the properties of the analysis techniques govern the minimum reconstruction error achievable. Hence, the methods described in Section 4 were designed not only to improve signal analysis, but also to eliminate reconstruction errors in the unquantised case, and minimise them when parameters are quantised.

In WMWI analysis, the pitch track used to warp the residual is optimised to ensure that the pulse peaks of critically-sampled pitch periods are aligned. No magnitude or phase information of the input speech is destroyed during the analysis process, hence, given an identical pitch track, an unwarping procedure can near-perfectly reconstruct the input signal. (Note that the only source of error in unquantised WMWI is due to the minor filtering errors of the warping/unwarping process.) Good speech quality can still be achieved if the synthesis pitch track is not exactly the same as the one used in analysis, but if the two tracks differ significantly, distortions may result. These are caused when the unwarping (or compression) procedure time-scales the warped samples by a series of factors dissimilar to the inverse of the original warping factors. The proposed pitch quantisation/

reconstruction technique has been designed to minimise the possibility of such an occurrence. However, in the event that errors do occur in the pitch estimates, precautions are taken to ensure that the loss of time-synchrony will only be temporary, and it will be immediately regained in the following frame.

5.3. Pitch Track Quantisation

The fundamental frequency, or pitch period, of voiced speech is of major importance in speech coding and accurate matching of the level of periodicity is crucial for good quality reconstructed speech. In this section, previous methods to quantise the pitch track are reviewed, and the significant attributes of the WMWI pitch track that need to be maintained, are identified. The details selected for transmission enable correct positioning of pitch pulses at the decoder, and a method to construct an accurate representation of the analysis pitch track from the transmitted parameters is described in Section 5.4.

5.3.1. Review of Pitch Quantisation Methods

The methods used to code the pitch parameter vary significantly between parametric coders and waveform coders. In parametric coders, pitch quantisation is usually orthogonal to the quantisation of other parameters. This enables efficient pitch quantisation methods to be employed, such as that proposed by Eriksson and Kang [Erik99b], which will not deteriorate the quantised quality of the waveform shapes. This approach is based on the quantisation of differential and logarithmic pitch values, requiring only 4 bits per pitch sample. Conversely, quantisation of the pitch lag in waveform coders such as CELP coders, affects the quantisation of the

excitation subframes. As a result, in many LPAS coders, the pitch predictor consumes a large proportion of the overall bit rate to represent several pitch updates per frame. Speech quality can be further improved by adopting high-resolution or fractional-pitch methods [Marq90][Kroo90] that refine the resolution of the pitch period search to a fraction of a sample by means of interpolation.

Also, methods which incorporate pulse position coding, as opposed to pitch coding, have been proposed. These methods aim to correctly position pitch pulses and minimise perceptual distortions, especially during transitional regions of speech [Sohn99][Stac98]. Some techniques used in low-rate coding to transmit pulse position information include the use of 34-bit block codes for encoding pitch markers (Single-pulse excitation models [Gran91]), or 8-bit look-up tables to represent the approximate position of one pulse per 20ms frame, plus the number of pulses between the current and previously transmitted pulse positions [Stac98].

The proposed WMWI pitch quantisation method aims to transmit pulse position information within a varying number of parameters, providing greater accuracy than the method of [Stac98], yet with higher efficiency than that of [Gran91].

5.3.2. Pitch Track Attributes

The phase of a CW is composed of two components: offset information, in the form of a linear phase component, and shape information. In the case of voiced speech, the phase offset effectively dictates the position of the pulse peak. In standard WI, this component of phase information is destroyed when CWs are cyclically rotated, during the transformation of the residual signal into a 2-D representation, to achieve

alignment. No memory of the original orientation of the CW is retained. By optimising the pitch track to ensure that all critically-sampled pitch periods of the warped signal are aligned (such that no rotational alignment is required), phase offset (or pulse position) information is integrated within the pitch parameter. Therefore, an accurate representation of the pitch track at the decoder corresponds to knowledge of the pulse positions (or, in the case of unvoiced speech, nominal period boundaries). This allows the pitch pulses of the reconstructed speech signal to be time-synchronised with those of the input signal.

If the analysis pitch track of WMWI, derived in Section 4.4, were to be perfectly recreated in the decoder, every pitch pulse position would need to be transmitted, requiring a large number of bits. Such accuracy is not essential to achieve high perceptual quality synthesised speech and waveform matching. It is necessary, however, to preserve the significant facets of the pitch track.

5.3.3. WMWI Pitch Parameters

In the pitch quantisation approach of WMWI, only one pitch value, τ_{curr}, is transmitted per frame; this could simply be the pitch of the final period of the frame, or alternatively, it could be a value adaptively selected to minimise the pitch track reconstruction error. The pitch period value is quantised to the nearest half-sample to provide increased resolution and reduce the accumulation of rounding errors when integer pitch is used. However, rather than simply interpolating between pitch updates as in standard WI, additional information is transmitted to allow accurate recreation of the pitch track. A block diagram showing the extraction of

the required pitch parameters is in Figure 5.3. The side information required is:

a) The number of pulse peaks in the frame, $N_p - 1$

The variable N_p includes the first pulse of the future frame, as the position of this peak is required to determine the pitch of the last portion of samples in the current frame. Hence $N_p - 1$ represents pulses in the current frame only.

b) The number of whole periods in the frame which do not contain a pulse, N_u

This is required during unpulsed-to-pulsed transitions only and is in addition to Parameter a. Since any value of pitch may be assigned to the initial samples of the frame to ensure correct positioning of the first pulse, extra information is required to ensure this pitch value can be correctly estimated at the decoder.

c) Pulsed/Unpulsed Classification

This is mainly based on whether a stream of pulses has been detected or not, but also takes into account their locations in the frame and the classification of the previous frame. Since the analysis pitch track is defined for four possible frame transitions, knowledge of this parameter at the decoder is required to reproduce the structure of the pitch track.

d) Period Boundary Information, x

This is the most significant parameter for ensuring the waveform-matching objective is met. The number of *warped* samples at the end of the frame that do not make up a whole pitch period, x, are transmitted to ensure that the input and output streams are in synchronisation at the beginning of every frame. This is equivalent to transmitting the warped position of the first pulse of the future frame (or last pulse of the current frame).

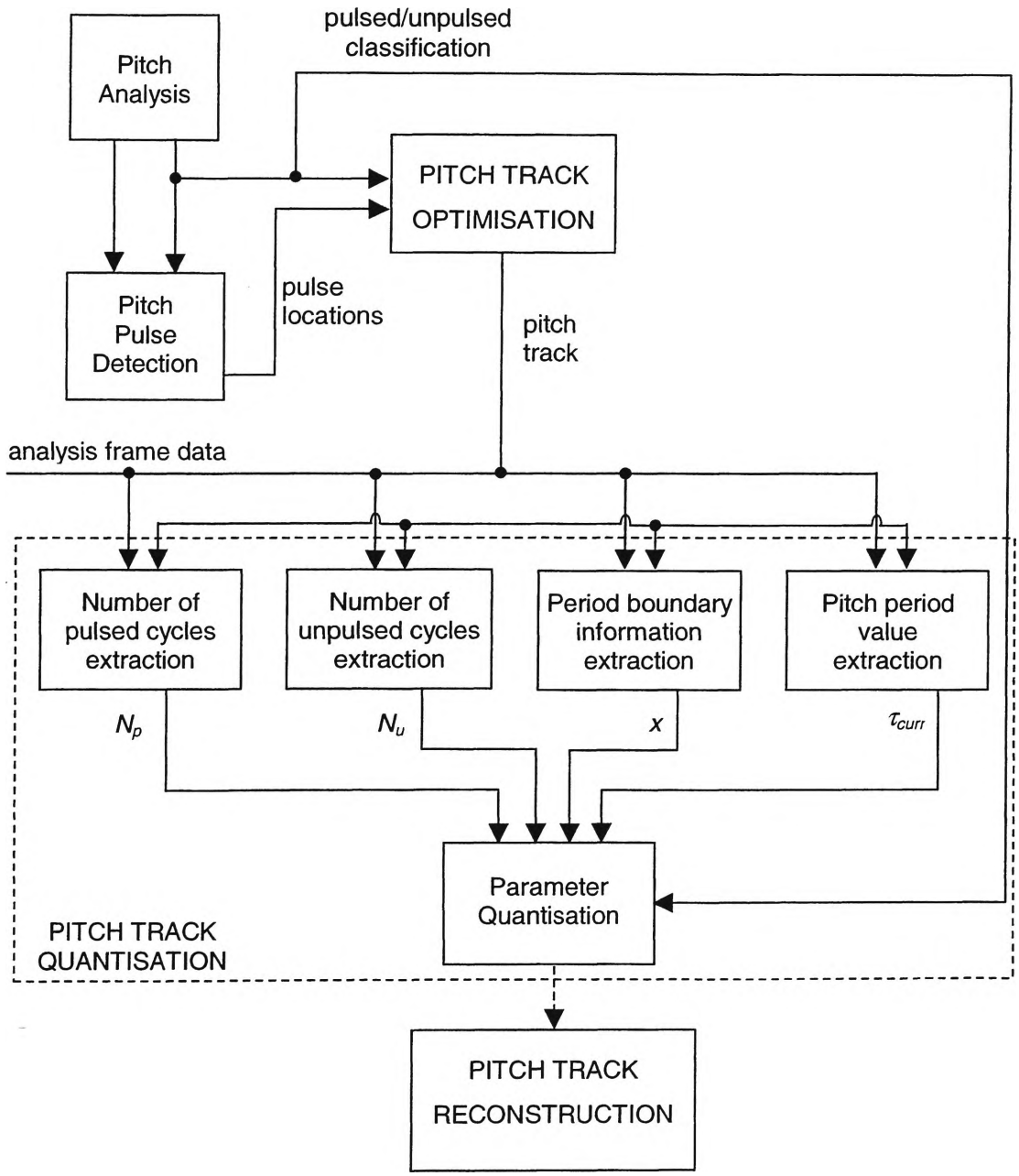


Figure 5.3. Block diagram of apparatus used to quantise the pitch track to enable time-synchronous signal reconstruction

In the event that a channel fades and frames of data are lost, the algorithm rectifies itself within one frame (mainly due to the consistent knowledge of the boundary parameter, x). The above pitch side information specifies the configuration of periods within the frame, maximising the accuracy of the pitch track reconstruction. In the case of strongly voiced or unvoiced speech, not all parameters are significant. However, if the pitch has rapid variation, as transition regions often do, to avoid pitch errors, all five parameters (a single pitch value plus 4 additional factors) are necessary to ensure the pitch track is accurately reconstructed. Note that while this information requirement may seem high, phase offset information is preserved, allowing favourable unwarping of the signal and time-synchronous reconstruction which facilitates waveform matching. Hence, the bit requirements, tabled in the following sections, must be compared with waveform coders, such as CELP, ADPCM, etc. In comparison to these coders, the bit allocation is not unreasonable.

5.4. Pitch Track Reconstruction

Given the quantised pitch parameters, the pitch track is reconstructed at the decoder. In Section 4.4, the pitch track was defined for a set of four possible frame types: Continuously Pulsed, Continuously Unpulsed, Unpulsed-to-Pulsed, and Pulsed-to-Unpulsed. This same classification is used in the reconstruction, and each case is described in detail below. In accordance with the analysis pitch track, all samples within a pitch period carry the same pitch value. To reveal the advantages of this technique, speech synthesised using the proposed pitch track reconstruction

is compared to speech synthesised with a pitch track recreated using standard WI methods.

5.4.1. Continuously Pulsed Frame

One pitch period value is transmitted every frame, and this value may be optimally selected. For the sake of simplicity, the transmitted pitch value, τ_{curr}, is chosen to be the pitch of the period overlapping the boundary between the current and future frames. Hence, given τ_{curr}, and the pitch of the corresponding period of the previous frame, τ_{prev}, the pitch of the intermediate periods is chosen to satisfy the constraints specified by the side information. Therefore, in a frame of F samples, there must be N_p-2 whole pitch periods, and one pitch period overlapping each of the frame boundaries. The extent of the overlap of a period from one frame to the next is expressed in the parameter x. While the CWs are quantised with their greatest energy centred in the waveform to minimise discontinuities, for the purposes of deriving a pitch track to obtain pitch normalisation, the assumed period boundaries are shifted to occur at the pulse positions, rather than halfway in-between the pitch pulses. Hence, x is adjusted to reflect the shift in the assumed period boundaries by half a period as follows:

$$x_{adj} = \begin{cases} x + \dfrac{L}{2}, & x < \dfrac{L}{2} \\ x - \dfrac{L}{2}, & \textit{otherwise} \end{cases} \tag{5.1}$$

A simple diagram of the frame in the warped time domain is shown in Figure 5.4.

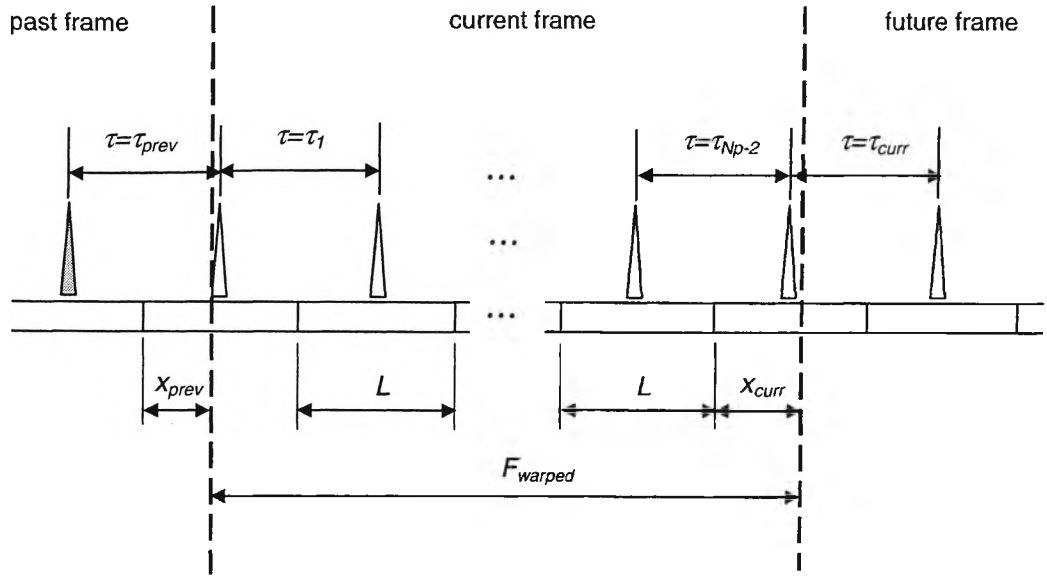


Figure 5.4. Diagram of pitch periods within a "Continuously Pulsed" frame in the warped time domain.

The pitch of the m^{th} intermediate period, τ_{p_m}, is searched subject to the following constraints:

$$\sum_{m=1}^{N_p-2} \tau_{p_m} = F - \left(\frac{L - x_{prev,adj}}{L}\right)\hat{\tau}_{prev} + \frac{x_{curr,adj}}{L}\hat{\tau}_{curr} \tag{5.2}$$

where

$$\tau_{p_m} = \frac{(N_p - 1 - m)\hat{\tau}_{prev} + m\hat{\tau}_{curr}}{N_p - 1}, \qquad 0 \le m \le N_p - 1 \tag{5.3}$$

Subscripts *prev* and *curr* relate to the previous and current frames respectively and ^ relates to quantised values.

The bit allocations required for the pitch are shown in Table 5.1. A warped pitch period length, *L*, of 256 samples, and a frame length, *F*, of 200 samples (25ms at a sampling rate of 8kHz) is used.

Table 5.1. Bit allocation required for accurate pitch reconstruction of Continuously Pulsed frame

	bits / samples	**Frequency (Hz)**	**Rate (bits/sec)**
Pitch	8 / 200	40	320
Number of periods containing a pulse	4 / 200	40	160
Pulsed/Unpulsed Classification	1 / 200	40	40
Period Boundary Information	8 / 200	40	320
TOTAL	**21 / 200**		**840**

5.4.2. Continuously Unpulsed Frame

For frames in which no pulses are detected, only the pulsed/unpulsed classification needs to be transmitted, as shown in Table 5.2. The pitch is fixed at a nominal, predetermined value for the entire frame. The lack of pitch information which needs to be represented in continuously unpulsed frames allows surplus bits to be used for error robustness and the transmission of additional information for future transitional frames.

Table 5.2. Bit allocation required for accurate pitch reconstruction of Continuously Unpulsed frame

	bits / samples	**Frequency (Hz)**	**Rate (bits/sec)**
Pulsed/Unpulsed Classification	1 / 200	40	40
Additional Information	20 / 200	40	800
TOTAL	**21 / 200**		**840**

5.4.3. Unpulsed-to-Pulsed Transition

During analysis, the pitch during the noisy part of an Unpulsed-to-Pulsed frame was chosen such that the first pulse peak was warped to the desired pulse position. The proposed reconstruction technique aims to reproduce this in the decoder. This requires more side information to be transmitted than for the Continuously Pulsed case. To improve accuracy, and accommodate pitch variations at the beginning of pulsed regions better, an additional pitch, the pitch of the first pulsed period, τ_0, is also transmitted. Since continuously unpulsed frames require only the pulsed/unpulsed classification to be sent, the delay of the SEW/REW decomposition allows extra information to be transmitted for the transitional frame with information for the previous unpulsed frame. This requires the restriction that an Unpulsed-to-Pulsed frame must follow a Continuously Unpulsed frame. The bit

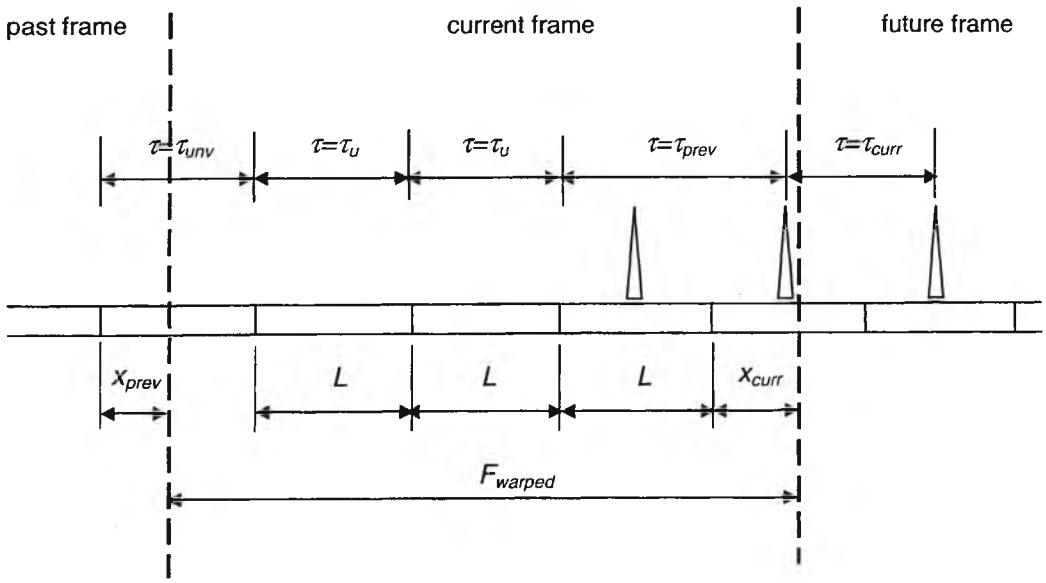


Figure 5.5. Diagram of pitch periods within a "Unpulsed-to-Pulsed" transition frame in the warped time domain.

Table 5.3. Bit allocations required for the previous and current frames to obtain accurate pitch reconstruction of an Unpulsed-to-Pulsed frame

	bits / samples	**Frequency (Hz)**	**Rate (bits/sec)**
For current frame			
Pulsed/Unpulsed Classification	1 / 200	40	40
For future frame			
Pitch	8 / 200	40	320
Number of periods containing no pulse	3 / 200	40	120
Additional Information	9 / 200	40	360
TOTAL	**21 / 200**		**840**

(a) Previous Continuously Unpulsed frame bit allocation

	bits / samples	**Frequency (Hz)**	**Rate (bits/sec)**
Pitch	8 / 200	40	320
Number of periods containing a pulse	4 / 200	40	160
Pulsed/Unpulsed Classification	1 / 200	40	40
Period Boundary Information	8 / 200	40	320
TOTAL	**21 / 200**		**840**

(b) Current Unpulsed-to-Pulsed frame bit allocation

allocation is detailed in Table 5.3. The pitch of the k periods containing no pulse, τ_u, is calculated as follows:

$$k\tau_u = F - \frac{\sum\left(L - x_{prev}\right)}{\sum L}\hat{\tau}_{unv} + \frac{3}{2}\tau_0 + \sum_{m=1}^{N_p-3}\tau_{p_m} + \frac{x_{curr,adj}}{L}\hat{\tau}_{curr}\frac{\sum}{\sum} \tag{5.4}$$

where τ_{unv} is the nominal pitch for periods containing no pulse, and τ_{p_m} is the pitch of the periods containing a pulse, as expressed in Equation (5.3), where $\tau_{prev} = \tau_0$. The transitional frame with its associated pitch period values is shown in Figure 5.5.

5.4.4. Pulsed-to-Unpulsed Transition

For frames in which a train of pitch pulses decays and unvoiced speech begins, the transmitted pitch value, τ_{curr}, is the distance between the last two detected pulses of the frame. The pitch of the intermediate pulsed periods, can then be found by Equation (5.3). The pitch of the remainder of the frame is set at the predetermined unvoiced value. A diagram depicting the pulsed-unpulsed transition is in Figure 5.6, where n_p=4. The bit allocation required is identical to that of Continuously Pulsed frames shown in Table 5.1.

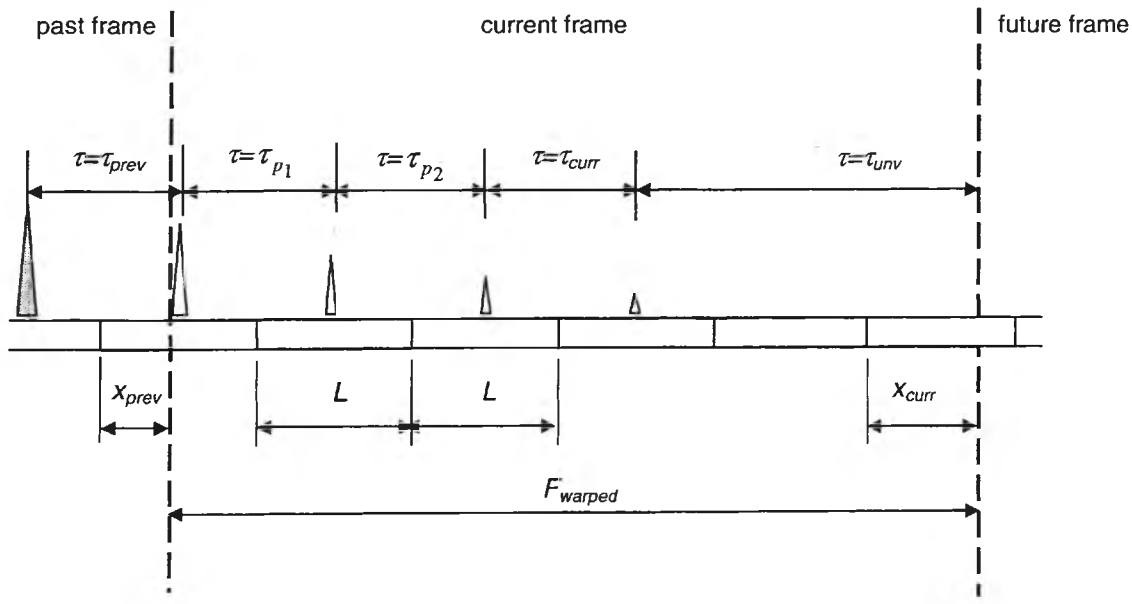


Figure 5.6. Diagram of pitch periods within a "Pulsed-to-Unpulsed" transition frame in the warped time domain.

5.4.5. Comparison of WMWI Modeling and WI Modeling

The WMWI reconstruction method allows the opportunity for very high quality coded speech, based on waveform matching. With the unquantised parameter set, near perfect reconstruction can be achieved, while at coding rates around 4kbit/s and above, good quality, time-synchronous output speech can be produced. Alternatively, an approximate reconstruction technique can be applied if the transmission rate is required to be reduced even further. This approximate method is used in standard WI synthesis and employs interpolation between the CWs. The combination of this technique with the WMWI front-end allows a much more detailed pitch-synchronous signal analysis than that of standard WI, while maintaining a low bit rate of around 2.4kbit/s. Since waveform matching is compromised, the phase offset of the pitch pulses is insignificant and therefore, the pitch track can be far less accurate.

To illustrate the waveform matching quality of WMWI synthesis as opposed to standard WI synthesis, a section of residual was analysed using WMWI and then reconstructed by each of the two techniques (Figure 5.7). In both cases, two SEWs and four REWs were transmitted per frame. It can be seen that the pitch pulses of the signal reconstructed using the WMWI technique (Figure 5.7b) are well aligned with those pulses of the input signal (Figure 5.7a), allowing waveform matching. On the other hand, while the output signal formed by the modeling of standard WI is a good representation of the input, the significant waveform features are offset from their original positions (Figure 5.7c). This may result in an extra

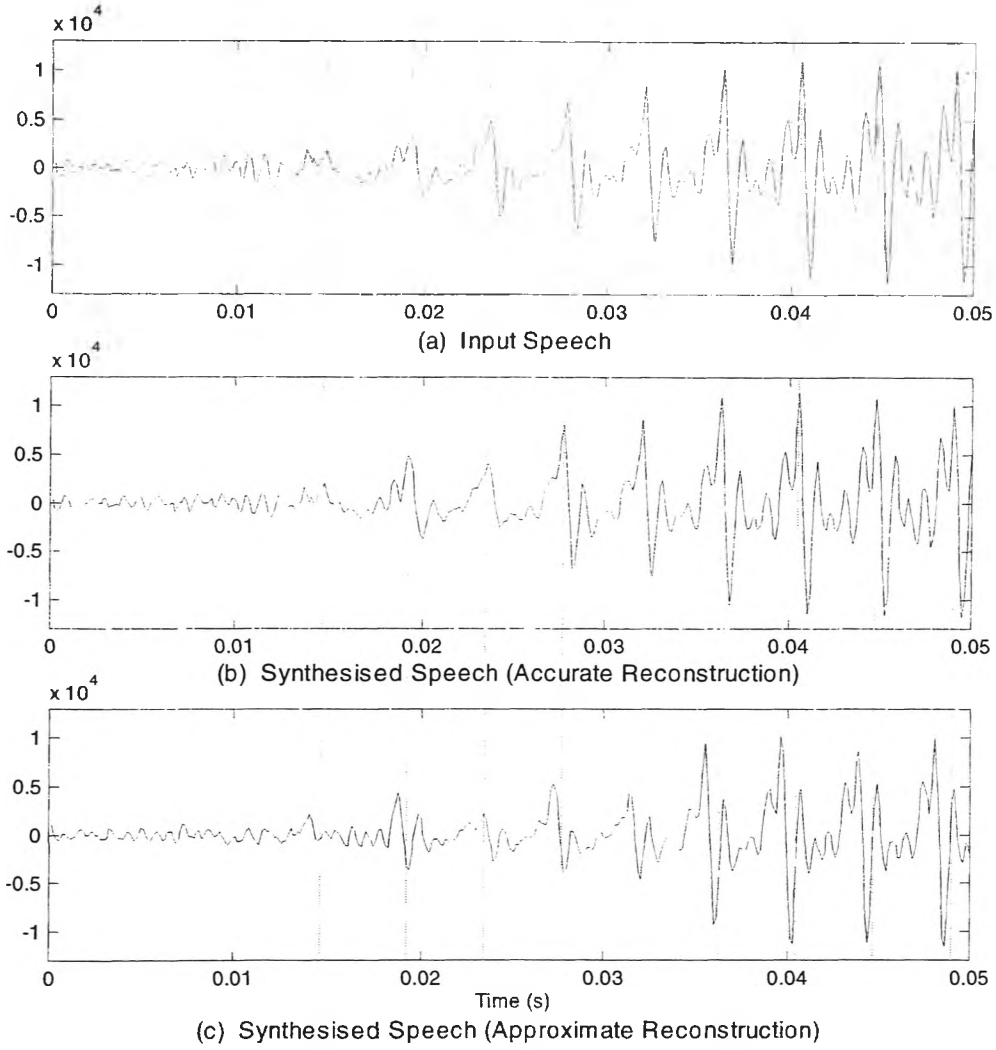


Figure 5.7. Comparison of the accurate reconstruction method and approximate reconstruction method

period being gained (or one omitted) in the reconstructed signal, leading to audible distortions. While the asynchrony of the pitch pulses of Figure 5.7c is a very apparent point of dissimilarity between the accurately and approximately reconstructed sequences, this is perhaps not the most significant difference. An advantage of the WMWI synthesis is the amount of detail of the input signal maintained during reconstruction, as opposed to the interpolated detail of the WI synthesis approach. This allows the preservation of, for example, pitch period lengths which do not vary at a consistent rate. Also, it can be observed that the detail in the transitional region from unpulsed to pulsed speech is closely matched with the accurate reconstruction (Figure 5.7b), but not with the approximate reconstruction (Figure 5.7c).

5.5. Quantisation of the Decomposed Surfaces

The nature of voiced and unvoiced speech differ significantly, both in their rate of evolution and perceptually significant content. This motivated the use of a decomposition to separate these components to allow each to be quantised with an accuracy based on their perceptual significance. In this section, various techniques to efficiently quantise and transmit the decomposed components, the SEW and REW, across a channel (Figure 5.8) are discussed. The merits and disadvantages of coding each parameter in the warped time domain, unwarped time-domain and frequency domain are described.

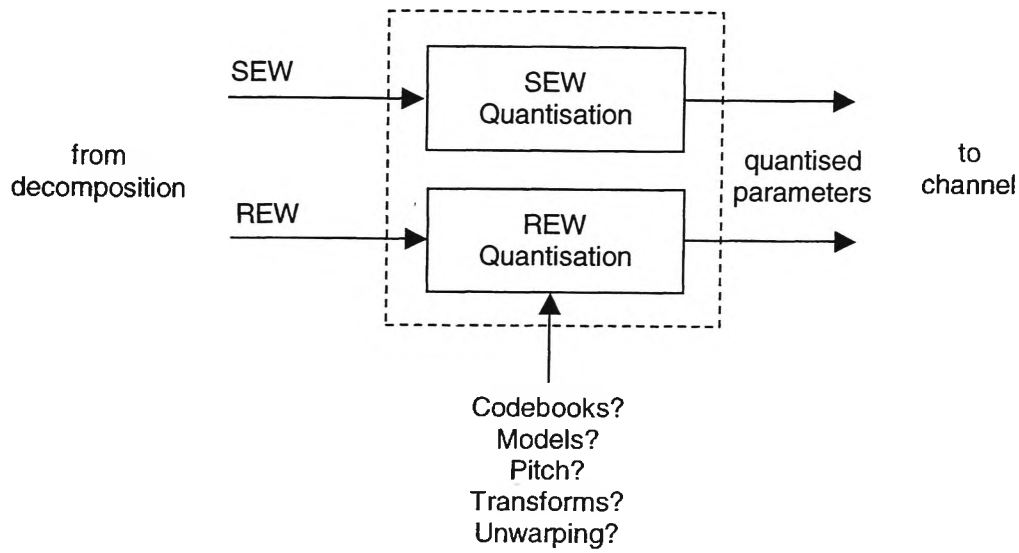


Figure 5.8. Quantisation of the decomposed components

Two fundamental decisions are:

1. *Warped Domain or Unwarped Domain Quantisation?*

   The purpose of warping is to normalise the pitch variations of the speech signal to enforce a constant length pitch period. This allows an accurate representation of speech evolution and good decomposition. Hence, warping yields benefits to sections of speech containing periodicity. The periodic nature of the SEW, and its common structure with a central pulse in each pitch cycle during voiced speech, make it advantageous to quantise the SEW in the warped time domain. However, unlike the SEW, the REW has no underlying periodicity, nor any standard shape. It therefore suggests that the REW may not benefit at all by the warped domain representation. On the other hand, pitch-synchronous quantisation of the REW enables the easy application of operations such as time-domain masking or perceptual weighting due to the fixed position of the high-

energy pulse regions of the SEW, and allows the quantised SEW and REW signals to be well-synchronised.

2. *Time Domain or Frequency Domain Quantisation?*

   Due to the warping procedure, which removes the pitch variations of the input speech signal, time domain quantisation can be performed using (fixed-length) VQ techniques. Optimisation of the pitch track, such that significant features of each pitch period are aligned, makes this technique effective.

   However, due to the pitch fluctuations of speech, quantisation in the frequency domain is best performed using an adaptation of VDVQ. As described in Section 4.3.3, while the spectral transformation of the warped pitch cycles results in a constant number of transform coefficients, the number of significant transform coefficients for each CW is approximately equal to the *pitch period* for real-valued transforms. For complex-valued transforms, only the magnitudes of $\frac{pitch\ period}{2}$ coefficients are significant. Hence, the most effective VQ technique is one that considers only these components. This results in the need to quantise variable length sequences, similar to those obtained if the transformation was performed on unwarped, pitch-length CWs.

5.5.1. SEW Quantisation

The low evolution bandwidth of the slowly evolving waveform allows a low update rate, and achieves good perceptual quality with interpolation of the waveforms [Klei95]. Hence, the SEW can be downsampled prior to transmission. It was found

that two SEW updates per frame produced a good representation of the SEW evolution for the WMWI 4kbit/s framework, although improved quality can be obtained with a higher update, and hence, transmission, rate. Both time domain and frequency domain quantisation techniques are discussed for the SEW component and the preferred technique is identified.

A. Time Domain Quantisation

To reiterate, the main advantage of time-warping, is to facilitate the application of fixed-length, pitch-synchronous parameter extraction methods and provide for improved analysis across waveforms with similar features. Since the SEW contains the periodic parts of the speech signal, it predominantly carries with it the benefits of time-warping. Hence, it is preferable to quantise the SEW parameter in the warped time domain, as opposed to the unwarped time domain, to take advantage of its constant periodicity and convenient 2-D surface representation which allows easy access to individual, aligned SEW pitch cycles.

Quantisation of the SEW parameter in the warped time domain is performed most efficiently using VQ techniques. Since the SEW vectors have a wide dynamic range, gain-shape VQ [Gers92] is used. The subjective quality of the reconstructed speech can be improved by incorporating perceptual weighting into the codebook search. This may be done by shaping the spectrum of the quantisation noise to have less energy in the regions of anti-formants. Alternatively, since pitch pulses are exactly aligned, it is easy to emphasise the part of the signal which is more perceptually

relevant. For voiced speech, this is the pulse area, including a small segment directly preceding the pulse.

Codebook Design

The WMWI analysis techniques ensure all pitch periods are warped to have a constant pitch, with the pulse peak occurring at the central sample within the pitch period. For female speakers (high pitch frequency), the warping factor is greater than that for male speakers (low pitch frequency). As a result, the pulses in the warped time domain have a range of durations (widths), depending on the pitch of the input speech which dictates the warping factor. This is illustrated by Figure 5.9, whereby the input pitch cycle is warped by two different factors, y_1 and y_2, where $y_2 > y_1$.

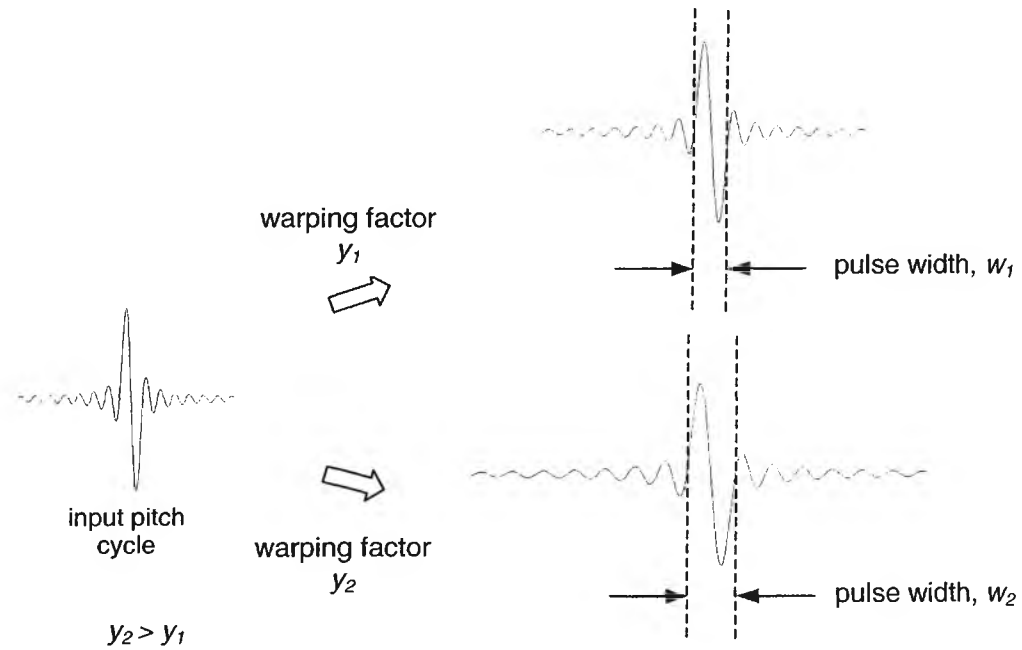


Figure 5.9. The warping of a pitch cycle by different factors

While the alignment of the *positive* pulse peak of each CW is enforced due to the pitch track formation algorithm, the result of warping with different interpolation factors is that the position of the *negative* pulse peak may vary greatly. This is shown in Figure 5.10. Assuming the pitch track to be constant for the whole pitch period, the warping factor for the first waveform is $\frac{L}{p_1}$, which is much greater than the warping factor for the second waveform, $\frac{L}{p_2}$. This causes the negative pulse peak of the first waveform to occur further along the warped time scale than that of the second waveform. If all the warped CWs were trained simultaneously, the

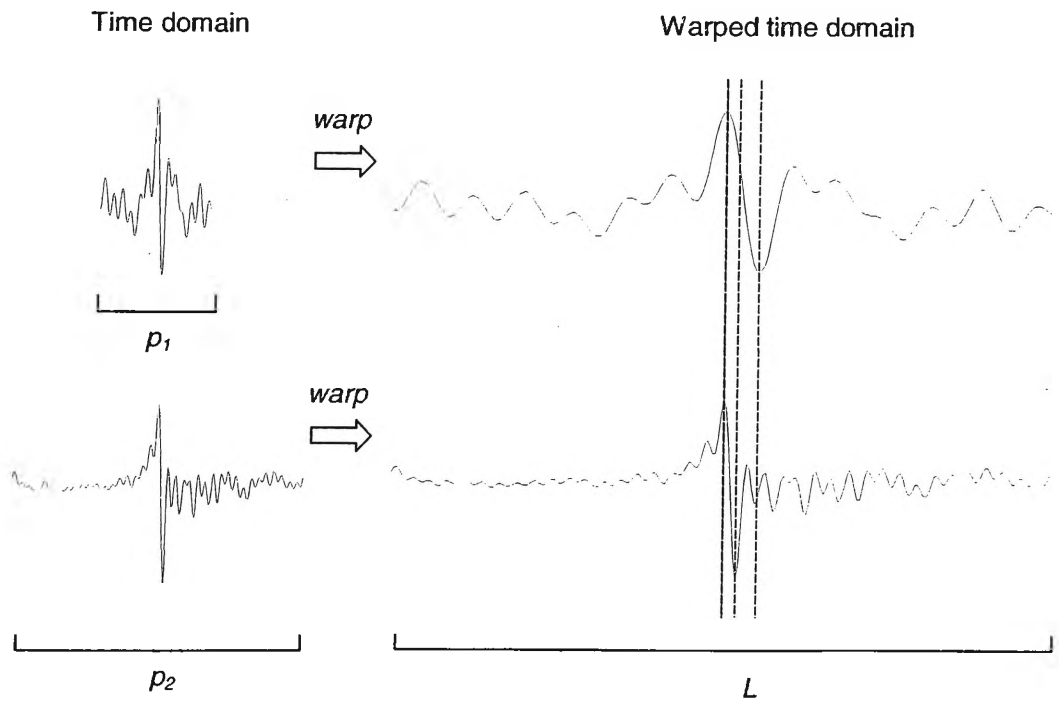


Figure 5.10. Warping of characteristic waveforms of different pitch values to a constant length, L. While the positive pulse peaks are aligned, the negative pulse peaks are offset.

negative pulse peaks of the training data set would not occur in the same warped location of each CW and thus, would not reinforce each other during the codebook design process. This results in codebook vectors with a large positive peak, but very small (if any) negative peak. Hence, a multi-codebook approach is adopted, in which the codebooks are pitch dependent. The range of possible pitches (20-147 samples) is divided into 10 sub-ranges, and a separate codebook is trained for each sub-range. Once the codebooks have been created, the pitch of the CW determines which codebook to search during quantisation. While more memory is required at the encoder and decoder to store the codebooks, no additional bits are necessary to specify the particular codebook to use. To minimise errors due to quantised pitch estimates falling within a different sub-range to that of the original pitch value, the codebooks are structured such that they contain similarly shaped vectors at the same index position.

The SEW gain is scalar quantised (SQ) on a logarithmic scale. Initially, two gain parameters were sent per frame: a 3-bit absolute gain for the first shape vector, and a 2-bit differential gain adjustment for the second shape vector. However, due to sharp increases in gain at the onsets of voiced speech, the differential gain approach could not adequately track the sudden gain changes in the SEW, and an absolute gain is transmitted with each shape vector. Techniques such as combined step-differential gain quantisation may provide improved representation of both large and small gain changes.

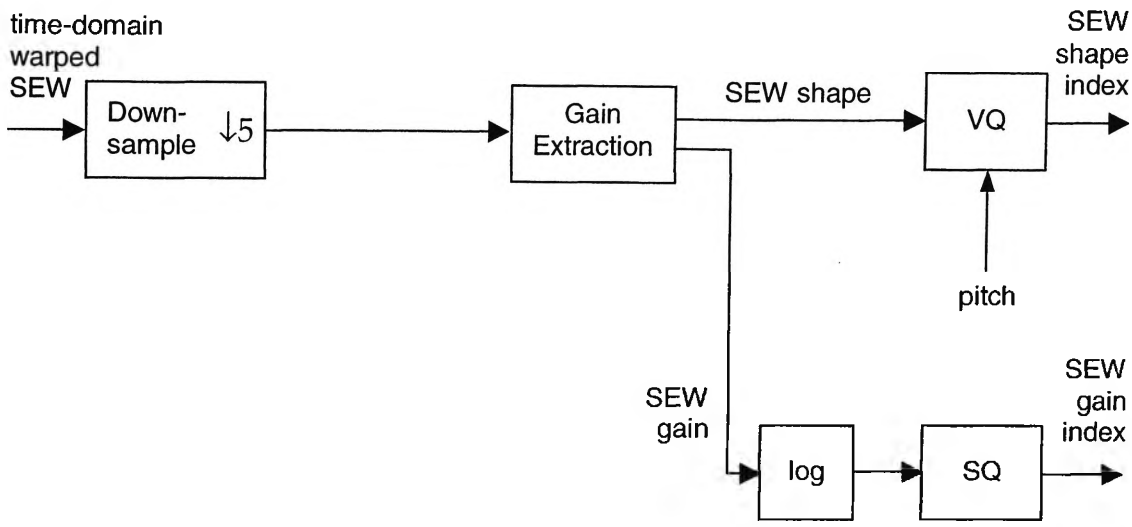


Figure 5.11. Quantisation of the SEW component in the warped time domain

Table 5.4. SEW bit allocation using shape-gain VQ of time-domain waveforms

	bits / samples	**Frequency (Hz)**	**Rate (bits/sec)**
SEW Shape	7 / 100	80	560
SEW Gain	3 / 100	80	240
TOTAL	**20 / 200**		**800**

The bit allocation for the SEW shape and gain is shown in Table 5.4 and a block diagram of the SEW quantisation in the time domain is given in Figure 5.11. Two SEWs are transmitted per frame, and the critically-sampled SEW surface is reconstructed at the decoder by interpolation of the SEW evolution.

B. Frequency Domain Quantisation

As an alternative, quantisation of the SEW in the frequency domain is also considered.

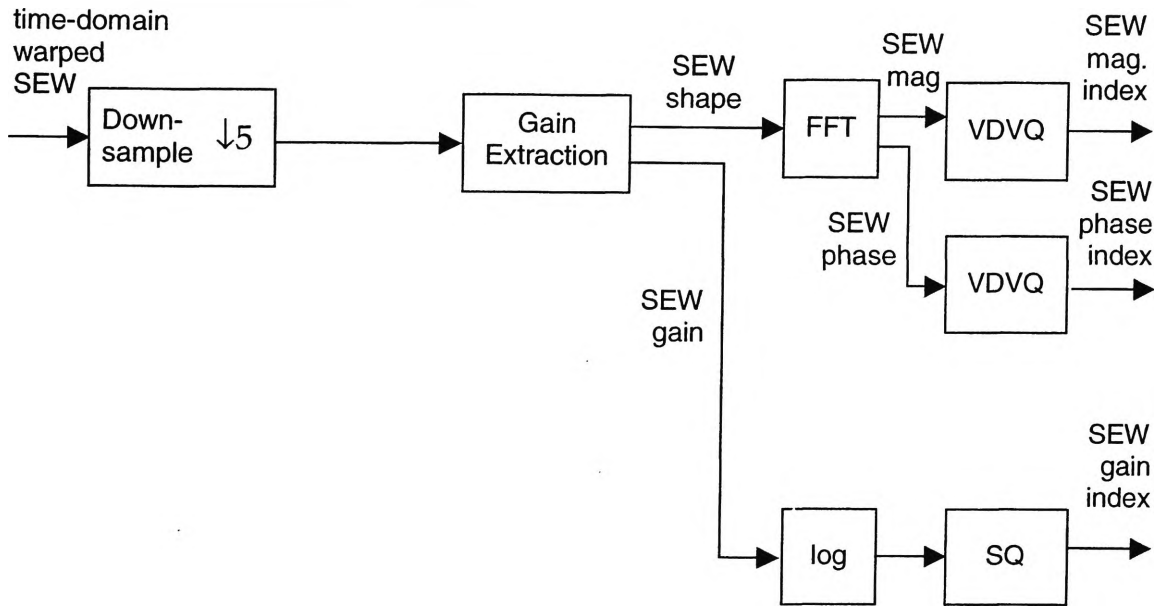


Figure 5.12. Quantisation of the SEW component in the warped frequency domain

Magnitude Quantisation

As described earlier in this section, quantisation of the transform coefficients is best performed using VDVQ, due to the effects of warping in the transform domain. The simple Fast Fourier Transform (FFT) was used to convert the signal to the discrete frequency domain, and VDVQ was applied to the magnitudes of truncated pitch-length coefficient series.

Phase Quantisation

VDVQ of phase can also be carried out; the main difference between phase and magnitude quantisation is that in the phase case, we must take into account phase wrapping. Phase quantisation is described in Section 3.9.7. Alternatively, phase models extracted from voiced speech could be applied.

C. Preferred SEW Quantisation Technique

The disadvantage of frequency domain quantisation techniques, is that they fail to capitalise on some of the beneficial features of the warped-domain representation of the SEW. For example:

1. Quantisation of SEW vectors is a variable dimension problem, resulting in some degradation in the representation of sequences, as opposed to fixed dimension quantisation.
2. The perfect alignment of pitch pulses is not exploited, as phase quantisation and modelling techniques require the linear phase component to be removed before quantisation for the most effective results.
3. If phase models are applied, this results in the loss of waveform matching for the very perceptually important periodic component of the waveform.

Hence, *time domain* quantisation of the *warped* SEW sequences (Figure 5.11) is adopted in the WMWI coder.

5.5.2. REW Quantisation

It is widely recognised that only the magnitude spectrum of the REW carries perceptually significant information, and the quantisation accuracy required for this magnitude spectrum is low [Kubi93]. This enables a high REW sampling rate, and thus, high time resolution, to be maintained without corresponding to large increases in the bit rate.

Studies have indicated that the audible characteristics of pitch-synchronously modulated noise is different for female and male speakers [Skog97]. For high-

pitched sounds, the auditory system is most sensitive to low frequency noise in the valleys between the harmonics in the spectral domain. For low-pitched sounds, the sensitivity is strongest for high frequency noise in the valleys between the pulse peaks in the time domain. These findings are consistent with the observations of varying length time domain pitch cycles. The high energy of the pulse area relates to a much smaller proportion of longer (low pitch frequency) cycles, providing less masking to intermediate samples. Similarly, spectral harmonics are spaced further apart for high frequency sounds, providing less masking for intermediate frequencies.

These considerations suggest a dual quantisation technique may be appropriate for the coding of the REW (noise) spectrum, which devotes higher importance to the level of periodicity for female speakers, but greater significance to the phase spectrum of male speakers. Indeed, the REW quantisation technique recently proposed by [Kang99] is of this nature and contains frequency dependency.

In this section, two possible quantisation techniques are analysed for the REW parameter. The associated advantages and disadvantages of each are discussed and the preferred quantisation method is determined.

A. Time Domain Quantisation

Unlike the SEW component, the REW component does not contain periodicity, and hence, does not benefit greatly from the warped domain representation. In fact, the warping of noisy sequences creates more work for the quantisers as it alters the statistics of the formerly Gaussian random variable. It is then preferable to quantise

the time-domain representation of the REW in the unwarped domain, where the signal is not influenced by varying interpolation factors. REW quantisation in the unwarped time domain is performed using analysis-by-synthesis CELP techniques. A mixed CELP/PWI implementation was originally proposed to separately code the voiced and unvoiced components [Klei93b][Burn93], but was thrown out as being unnecessarily complex. The CELP techniques are reinstated here to give waveform matching of the REW. Here, we divide the frame into four (or five) subframes, each containing 50 (or 40) samples. The subframes are then perceptually weighted, de-emphasising the frequency regions corresponding to the formant frequencies, as greater amounts of quantisation noise are masked in frequency bands where the speech has high energy. The weighting filter, $W(z)$, is derived from the short-term predictor filter, $A(z)$. The amount of de-emphasis is controlled by the factor γ.

$$W(z) = \frac{A(z)}{A\left(\frac{z}{\gamma}\right)} = \frac{1-\sum_{i=1}^{p} a_i z^{-i}}{1-\sum_{i=1}^{p} a_i \gamma^i z^{-i}} \qquad 0 \le \gamma \le 1 \tag{5.5}$$

A codebook of Gaussian vectors or a trained codebook is then searched. It was found that a Gaussian codebook produced output speech with a more natural sounding quality than that obtained with a trained codebook. The codevector which minimises the perceptually weighted error between the synthesised and original waveform is selected. The analysis-by-synthesis scheme for REW quantisation is depicted in Figure 5.13. A gain term is transmitted at half the rate of the waveform shapes, and is interpolated between updates.

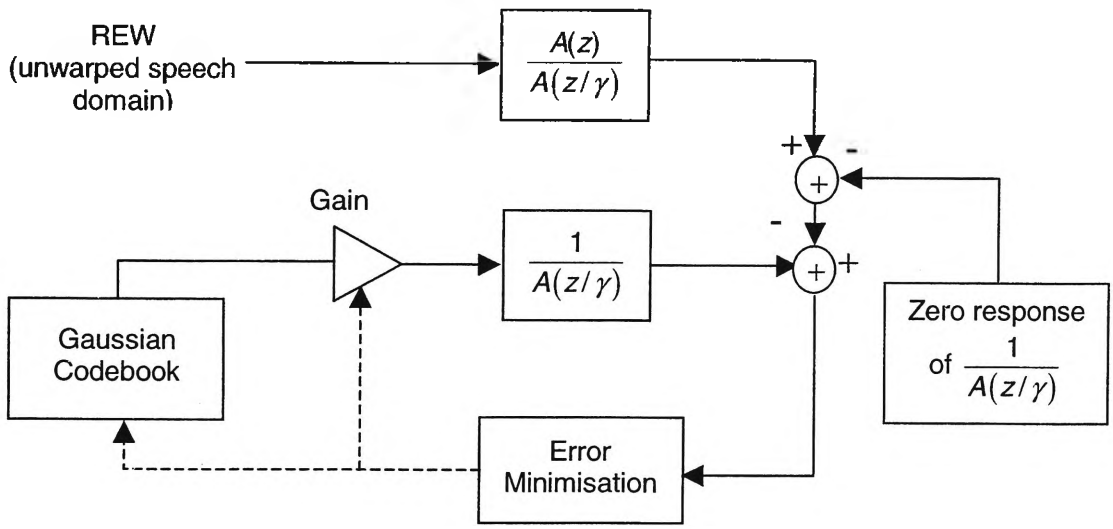


Figure 5.14. REW quantisation using the analysis-by-synthesis CELP architecture in the unwarped time domain

Table 5.5 outlines the allocation of bits used to quantise the REW waveform using four subframes is shown below. Increasing the number of subframes per frame can considerably improve the performance quality, and this facility may be used to achieve scalable coding.

Table 5.5. REW bit allocation for unwarped time domain approach

	bits/samples	**Frequency (Hz)**	**Rate (bits/sec)**
REW Shape	6 / 50	160	960
REW Gain	3 / 100	80	240
TOTAL	**30 / 200**		**1200**

Time-Domain Masking

During voiced speech, the most important segment of the period is the pitch pulse. It is essential that the REW, when combined with the SEW, does not destroy the

pulse shape. In addition, the region directly following the pulse is masked due to the high pulse energy, and hence, it has little influence on the perceptual speech quality. Experiments performed showed that if the central section of the warped time domain REW pitch period (corresponding to the pulse area of the SEW) was zeroed out for voiced segments only, the effect could not be detected in the output. Hence, this section of the REW cycle does not need to be considered during quantisation. This is easy to perform in the warped domain due to the consistent positioning of pitch pulses. In the unwarped domain, time-domain masking can still be applied by calculating the pitch pulse locations from the accurately reconstructed pitch track.

B. Frequency Domain Quantisation

Waveform matching of the REW, as performed in the time domain approach, does not seem practical due to the large degree of randomness and lack of pattern of the REW signal. While the proposed coder possesses, as one of its significant features, the ability to match the input and reconstructed speech signals, waveform matching is really only significant for the perceptually important characteristics of the signal. The accurate analysis of the CW evolution results in these characteristics being separated into the SEW component. Hence, while the loss of detailed phase information was undesirable for the quantisation of the SEW component, loss of phase information for the REW component is not viewed as compromising the true intentions of the waveform-matching objective.

While time domain REW quantisation was best performed in the unwarped domain to avoid varying interpolation factors, this characteristic is not problematic in the frequency domain. Warped domain magnitude/phase quantisation is equivalent to unwarped domain magnitude/phase quantisation. Both involve the same number of significant frequency components and the application of VDVQ, however, there are advantages for performing transformations to the frequency domain on the warped signal. Firstly, the pitch-synchronous extraction of pitch periods is simplistic. Secondly, only one unwarping procedure needs to be performed (on the combined SEW and REW signals at the decoder), rather than two individual unwarping operations (one at the encoder (REW) and one at the decoder (SEW)). Finally, and most importantly, since the signal has a constant pitch period length which is a power of 2, fast transforms can be applied. This provides substantial computational savings.

Therefore, for frequency domain quantisation of the warped REW, the REW magnitudes are transmitted four times per frame and interpolated. The phase spectrum is not transmitted at all. In the decoder, the phase applied is random in nature with a uniform distribution. A gain term is transmitted at half the rate of the REW magnitudes, and is interpolated between updates. Hence, rather than sending an overall gain, separate gain terms are transmitted for the SEW and REW components. It was found that separate gain terms provided better control over the ratio between the SEW and REW components, which is known to be very significant in terms of producing natural sounding speech. It also allows gain-shape VQ techniques to be applied to each component. The bit allocation for the REW is shown in Table 5.6.

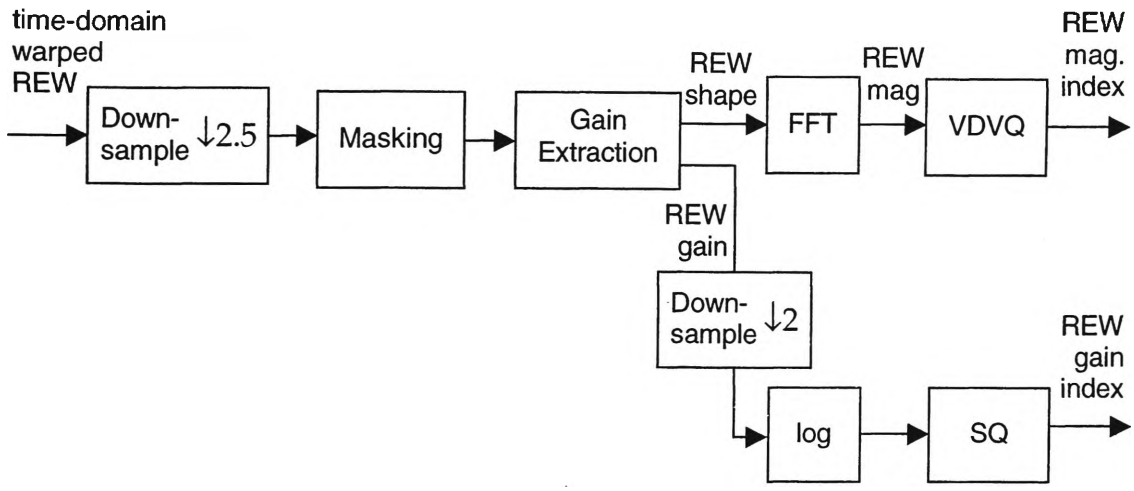


Figure 5.15. Quantisation of the REW component in the warped time domain

Table 5.6. REW bit allocation for frequency domain approach

	bits/samples	**Frequency (Hz)**	**Rate (bits/sec)**
REW Magnitude	6 / 50	160	960
REW Phase	-	-	-
REW Gain	3 / 100	80	240
TOTAL	**30 / 200**		**1200**

C. Preferred REW Quantisation Technique

Comparision of the time domain and frequency domain REW quantisation techniques illustrated the aforementioned audiblity characterisitcs of noise for female and male speakers. Informal listening tests showed that for high-pitched speech (pitch < 160Hz), the warped domain, magnitude/phase approach performed substantially better. For low-pitched speech (pitch ≥ 160Hz), the unwarped domain, waveform matching approach produced better quality than for high-pitched speech,

though the magnitude/phase quantisation was still marginally preferred in the majority of cases.

Hence, for the WMWI coder at around 4kbit/s, best results are obtained if the REW is quantised in the *frequency domain*, on the *warped* sequences. At higher bit rates, it is believed that advantages can be obtained using a dual quantisation mechanism, incorporating both time and frequency domain techniques, based on the pitch frequency. These higher rates will allow better waveform matching and a higher update rate of the time domain subframes.

5.6. WMWI Implementation at 4kbit/s

5.6.1. The WMWI Architecture

The analysis and decomposition process is shown in Figure 5.16. This includes the methods used to:

a) estimate and extract the spectral envelope from the speech signal using LP analysis (Section 2.5),

b) detect pitch pulses and form an optimised pitch track (Section 4.4),

c) time-warp the LP residual signal to have a constant pitch (Section 4.3),

d) critically-sample the warped signal to form an accurate representation of the LP residual evolution, and

e) resample and decompose the CW surface into perceptually different components to achieve higher coding efficiency (Section 4.6).

The novel pitch quantisation scheme has been discussed in detail in Section 5.3, while standard methods, e.g. multistage or split VQ, can be applied for the quantisation of the LSFs obtained from the LP parameters.

The preferred quantisation techniques for the decomposed SEW and REW components, as described in Section 5.5 are shown in Figure 5.17. These include:

a) down-sampling of the decomposed components to give 2 SEWs per frame and 4 REWs per frame,

b) gain normalisation of the SEW and REW vectors,

c) VQ of the SEW shape and SQ of the logarithm of the SEW gain,

d) application of the FFT to the REW vectors, and

e) VDVQ of the REW magnitude spectrum and SQ of the logarithm of the REW gain.

The reconstruction of the SEW and REW surfaces is depicted in Figure 5.18. This follows a direct reversal of the quantisation techniques, with the exception of the application of random noise to the less perceptually significant REW phase component.

Finally, the WMWI synthesis structure is shown in Figure 5.19, and is simply an inversion of the analysis structure. This gives the coder its waveform matching property. A description of the pitch track reconstruction scheme, the most detailed of the synthesis procedures, is provided earlier in this chapter in Section 5.4. The decomposed components are added together and resampled back to the critical-sampling rate using the reconstructed pitch track. The signal is then transformed to

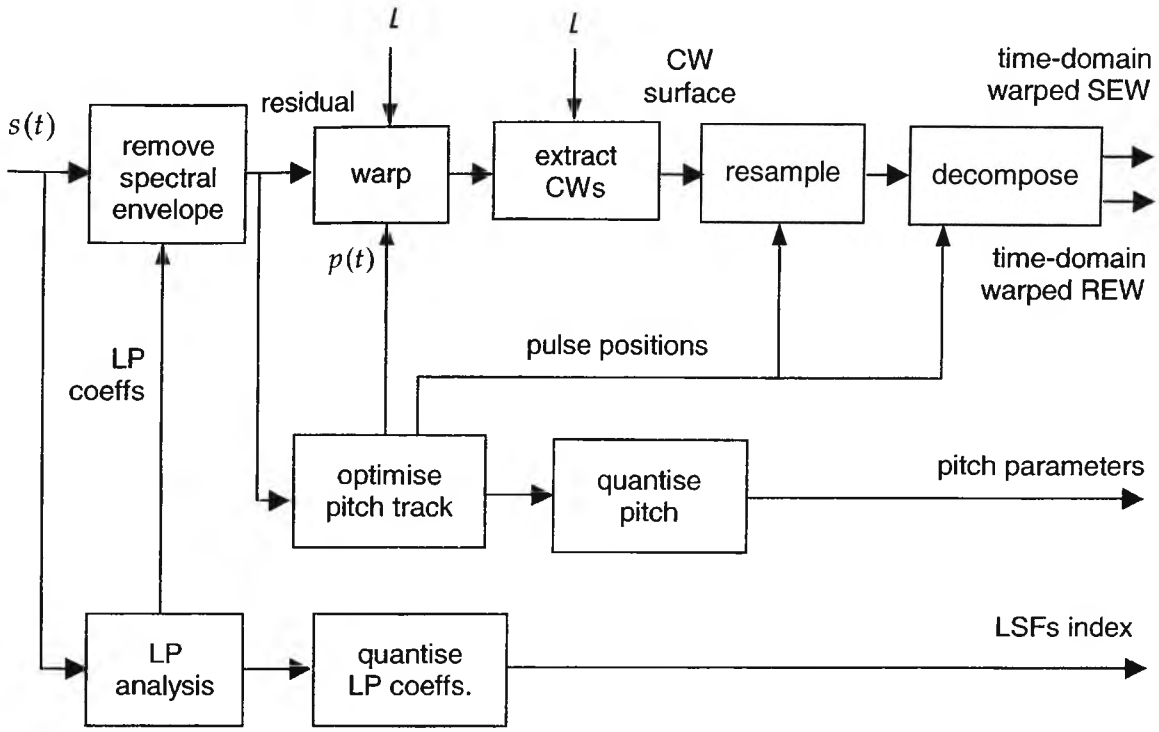


Figure 5.16. WMWI analysis architecture

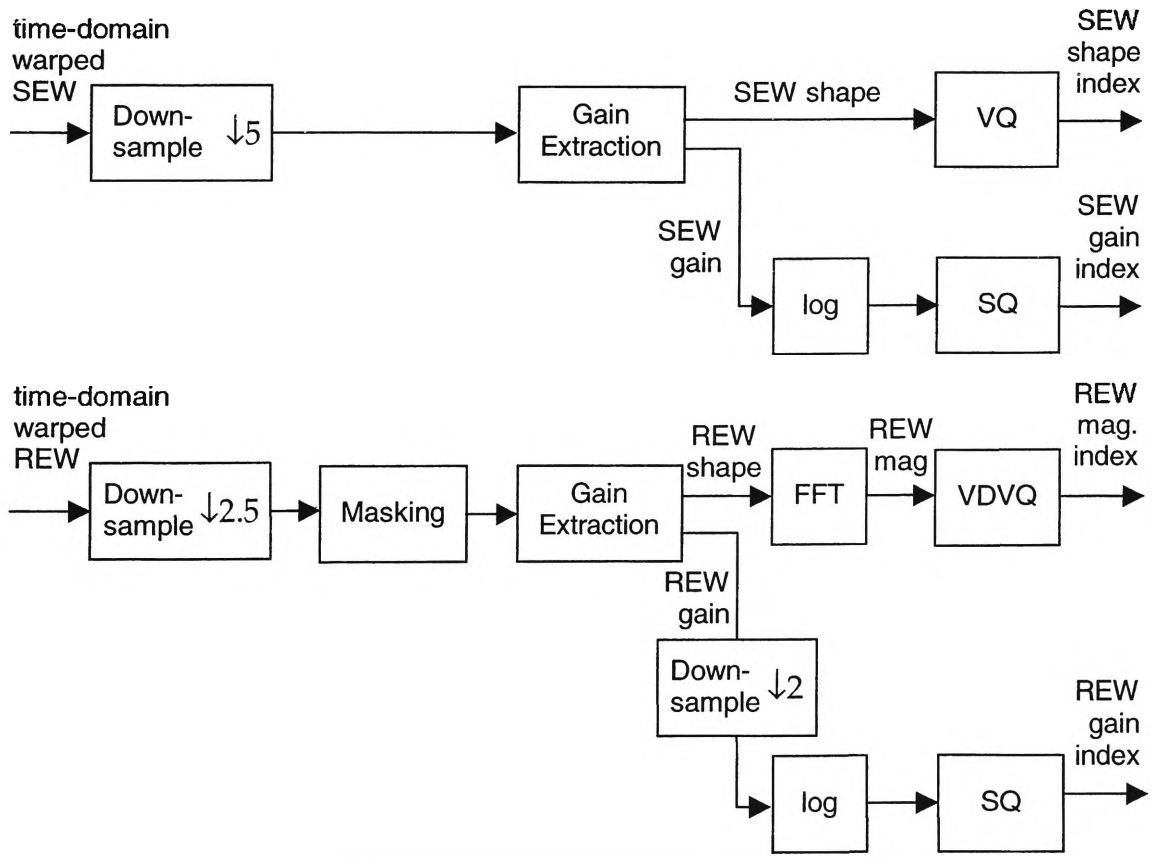


Figure 5.17. Quantisation of the SEW and REW parameters

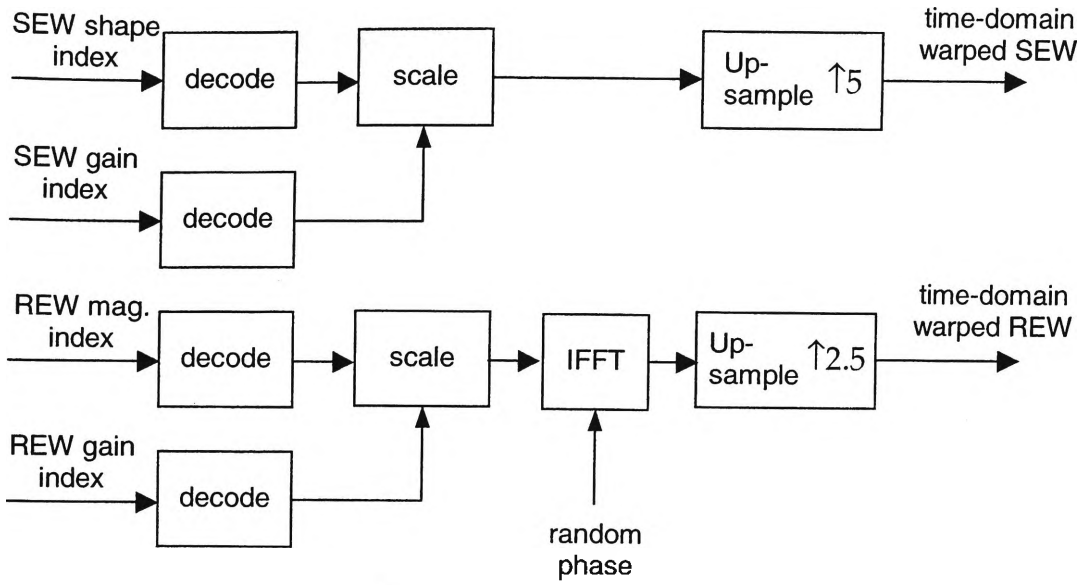


Figure 5.18. Reconstruction of SEW and REW components

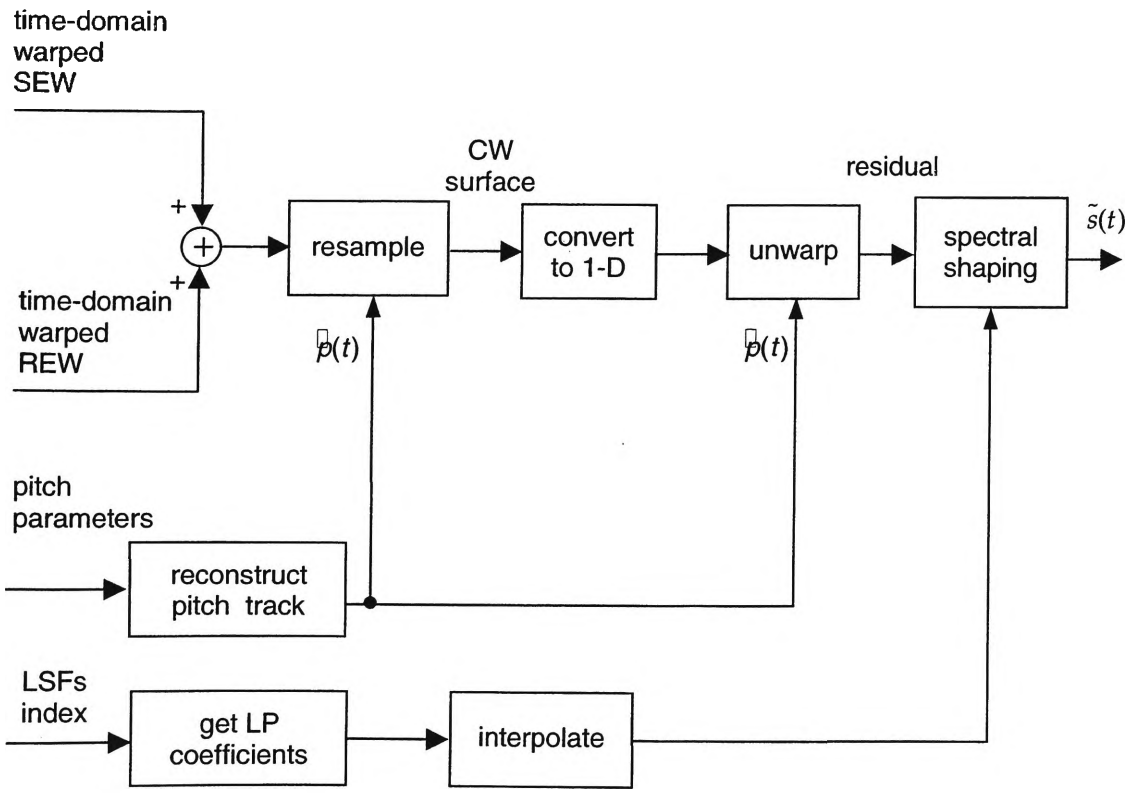


Figure 5.19. WMWI synthesis architecture

a one-dimensional representation and unwarped to produce the synthesised residual signal. Lastly, the spectral envelope is applied to generate the output speech. The conversion of the CW surface to one dimension during WMWI synthesis is much more straightforward than the conversion required in standard WI synthesis. No interpolation between pitch cycles is required, just the concatenation of CWs from the 2-D surface. Any edge effects caused by discontinuities between the endpoints of adjacent CWs are not detrimental, as they are smoothed out by the unwarping filter and hence, are undetectable in the unwarped residual signal.

5.6.2. Bit Allocation

The bit allocations for the WMWI implementation at 4kbit/s, using the quantisation techniques for the pitch and SEW/REW components described in this chapter are outlined in Table 5.7. Quantisation of the line spectral frequencies (LSFs), representing the LP coefficients which describe the spectral envelope, has not been covered in this thesis. Many techniques for efficient LSF quantisation have been presented in the literature, including predictive VQ, split VQ and multi-stage VQ (MSVQ) techniques [Pali93][Ohmu93][LeBl93][Pan98][Vasi99]. A total of 29 bits are reserved to represent the LSF parameter set; an allocation which will produce very good LSF representation at 4kbit/s. The aforementioned techniques are known to produce "transparent" quality (less than 1 dB spectral distortion (SD)) with this bit allocation and lower. For example, the tree-searched MSVQ technique of [Lebl93] achieves less than 1dB SD using 24 bits/frame, and the switched predictive MSVQ

Table 5.7. Bit allocation for the 4kbit/s WMWI coder

Parameter	bits/frame	Rate (bits/sec)
LSFs	29	1160
Pitch	21	840
SEW	20	800
REW	30	1200
TOTAL	**100**	**4000**

technique of the 1.7 kbit/s MELP coder [McCr98] achieves a frequency-weighted SD of less than 1 dB using only 21 bits/frame.

5.6.3. Subjective Performance

To measure the subjective quality of the WMWI quantised speech, a Mean Opinion Score (MOS) test was conducted, in which listeners were asked to grade the quality of each coded speech sentence on a scale of 1 (poor) to 5 (excellent). The WMWI coder was compared with the standard coders of G. 729 Annex C, an 8 kbit/s CS-ACELP coder [Sala98] and the 2.4kbit/s Federal Standard MELP coder [Supp97]. All WMWI parameters, except for the LSFs, were quantised and the test was performed for clean speech conditions. 30 listeners participated in the test, and 10 sentences (5 spoken by males) were coded by each speech coding algorithm. The resulting MOS scores are shown in Table 5.8.

The MOS scores obtained for the standard coders display differences to those published in the literature. This can be attributed to the large variation in output quality which pushes the G.729 scores higher, and the MELP scores lower. As a

Table 5.8. MOS scores for the coders: G.729, MELP, and WMWI.

	G.729	MELP	WMWI
MOS	4.08	2.94	3.28

Table 5.9. MOS scores for the coders: FS-1016, MELP and WMWI

	FS-1016	MELP	WMWI
MOS	3.22	3.33	3.53

second comparison, a similar MOS test was performed, but the G.729 coder was replaced with the 4.8kbit/s FS-1016 CELP coder [Camp89]. The new MOS scores shown in Table 5.9 provide a fairer measure of the output performance, as the coder qualities are more closely matched.

The results show that the WMWI coder performs better than both the FS-1016 coder and the MELP coder, achieving a MOS score of 3.5. While the output performance falls short of the desired benchmark of "toll-quality" standard, several improvements can be made. Since the coder produces near-perfect quality in the unquantised case, the need for further work is mainly in the area of SEW and REW representation (or in other decompositions). Also, no pre- or post-filtering was used for the WMWI coder, providing room for further perceptual improvement.

5.6.4. Delay

The delay of the WMWI coder is caused by 2 main operations:

a) the pitch pulse detection algorithm, and

b) the CW surface decomposition.

The pitch detection algorithm uses a composite correlation function with 5 segments, as described in Section 4.4, and requires a look-ahead of up to two frames (50ms) in its present implementation. The second frame of delay is required to accommodate signals with a pitch period greater than 100 samples, as the pitch track reconstruction scheme requires the location of the first pulse of the future analysis frame in order to assign pitch values for samples at the end of the frame. Alternative pitch arrangements may be formed for these infrequent cases, and hence the pitch look-ahead may be reduced to just one frame.

The delay introduced by the decomposition has been previously described in Section 4.6.5, and is equal to one frame, giving a total algorithmic delay of up to 75ms.

5.7. Summary

The WMWI reconstruction achieves waveform matching of the significant characteristics of the input speech signal by directly inverting the analysis procedures. This requires accurate transmission of the pitch track to maintain time-synchrony between the input and synthesised speech. A novel pitch quantisation and reconstruction technique is presented, in which significant facets of the pitch track are identified and transmitted as side information. This allows the locations of pitch pulse peaks to be accurately determined at the decoder. Alternatively, at lower bit rates, a low-resolution, interpolated pitch track may be applied, but this, however, compromises the waveform coding objective.

Several techniques for the quantisation of the decomposed waveforms of the LP residual are discussed. The main attributes of the warping procedure relate primarily to the pitch periodicity of voiced speech, enabling easy access to individual pitch cycles, precise knowledge of the pulse positions, and fixed length, constant rate analysis. In order to exploit the characteristics of the warped domain signal, quantisation of the SEW component is performed in the warped domain. Conversely, the REW component displays no periodic structure, and hence, the advantages of a warped domain representation lie in the ability to perform fast, constant length, transforms on consecutive pitch cycles, as well as easily apply time domain masking techniques.

In recent low rate coders, e.g. WI and sinusoidal coders, phase models are applied to the excitation signal or its components. While modelling of the frequency domain phase characteristics is not complimentary to the waveform matching property, the true objective of waveform matching relates to the perceptually significant speech components. Hence, the SEW component is quantised in the time domain using a multi-codebook VQ approach, in which all temporal variations (phase characteristics) are preserved. The REW component is best represented with the allotted bit allocation, in the frequency domain using VDVQ of REW magnitudes. At higher rates, it is foreseen that time domain REW quantisation techniques will be more prosperous.

The matching of the significant waveform features and perfect reconstruction property of the WMWI paradigm facilitates improved quality at higher rates, than

that obtained with completely parametric approaches, such as the standard WI coder. At 4kbit/s, the WMWI coder produces output speech with quality exceeding that of both the 2.4kbit/s MELP and 4.8kbit/s FS-1016 CELP coders.

Chapter 6

Conclusions and Further Research

The ability to quote is a serviceable substitute for wit.

-W. Somerset Maugham

6.1 Overview

As the use of cellular and satellite communication systems continues to escalate, so too does the expectation for improved service and higher quality transmission. This has seen an ongoing focus of research efforts into the area of speech coding, with the constant drive for lower bit rates without a corresponding decrease in output performance. While several speech coders have been standardised in recent years, there remains a void for good quality output speech at bit rates around 4kbit/s. Currently, the ITU-T is conducting standardisation efforts in this area. At this transmission rate, the limitations of both waveform coders, which produce excellent quality at higher rates, and parametric coders, which produce good quality at lower rates, become apparent, inhibiting their performance.

In this thesis, several techniques to bridge the performance gap between waveform coders and parametric coders, by integrating waveform matching properties into the Waveform Interpolation (WI) paradigm, are proposed. Specifically, the following issues are discussed:

- Multi-resolution decomposition of the speech evolution,
- Perfect reconstruction filter banks,
- Pitch-synchronous analysis,
- Time-domain warping,
- Optimised pitch track formation and quantisation,
- Quantisation of the (decomposed) excitation signal, and
- Time-synchrony and waveform matching.

This chapter highlights the major results obtained in each part of this thesis and also identifies areas for future research.

6.2 Decomposition Techniques

The decomposition method of standard WI has achieved a great deal of success for efficient quantisation. It is able to separate the very perceptually different slowly evolving components, relating to voiced speech, from noisy components which evolve at a much faster rate. This is achieved by simple linear filtering along the evolution of characteristic waveforms (CWs) and does not require a voiced/ unvoiced decision which was found to be a limitation in earlier parametric coders. This decomposition is taken a step further in Chapter 3, whereby the evolving waveform is decomposed into a series of multi-resolution surfaces, offering

advantages in terms of preferential bit allocation, based on the perceptual importance of the evolution frequency subbands, and the facility for feature extraction or noise suppression. The output comprises a periodic component in addition to the aperiodic component characterised at several scales. The proposed wavelet decomposition is implemented using a tree-structured perfect reconstruction filter bank, the design of which is flexible. The wavelet basis functions may be selected to match the characteristics of the signal and comply with the constraints (e.g. filter delay, rate of roll-off) on the decomposition. Several filter banks designs are discussed, with the preferred implementation being that of a low-delay FIR filter bank due to its simpler quantisation task compared to IIR filter banks with similar delay. The exact reconstruction nature and defined transmission rates of the decomposed surfaces provides for scalability in coding.

The following publications were produced as a result of this research: [Chon98a], [Chon98b], [Chon99a], [Chon00a].

6.3 Perfect Reconstruction Waveform Interpolation

6.3.1 Analysis and Decomposition

The WI coder is a parametric coder, limited at higher rates by its underlying speech production and perception model. Recently, interest has initiated the adaptation of the WI coder to incorporate the waveform coding or perfect reconstruction property. In Chapters 4 and 5, a Waveform-Matched Waveform Interpolation (WMWI) coder is developed. The proposed coder relies on the accurate detection of

pitch pulses and optimised formation of the pitch track and a reliable technique to perform this is presented in Chapter 4. The linear prediction residual signal is modified to have a constant pitch, and the pitch track creation technique ensures consistent positioning of pitch pulses at the centre of each warped pitch cycle, even after unvoiced segments. This allows for the effective decomposition of the evolving signal or efficient quantisation of the transformed sequences. Unlike previous WI coders, the analysis techniques of must ensure the preservation of all signal samples, both magnitude and phase characteristics. Several transformations and decompositions are discussed. Each may take advantage of the normalised pitch and consistent pitch cycle orientation of the residual. In the case of signal transforms, fast block transforms may be applied at a constant rate.

This work produced the following publications: [Chon99b], [Chon00b], [Chon00c], [Chon00e].

6.3.2 Quantisation and Reconstruction

The quantisation and reconstruction techniques applied to the WMWI parameters are presented in Chapter 5. Of primary importance is the desire to maintain the waveform coding objective. This translates, naturally, to good quantisation of the excitation signal in the form of decomposed components, but also to accurate representation of the pitch track, required to achieve time-synchrony between the input and synthesised speech signals. The significant facets of the pitch track which maximise the goal of correct pitch pulse positioning have been identified, and an efficient technique for quantisation and reconstruction of the pitch contour is

described. This avoids the audible artefacts which can occur in time-asynchronous coders when pitch cycles are omitted or repeated.

Quantisation of the decomposed surfaces may be performed by several methods and in several domains. It was determined that the slowly evolving component is best quantised in the warped time domain, since the primary objective of warping is to exploit pitch periodicity characterised by the SEW. On the other hand, the random nature of the REW results in reduced advantage of warping, however, the warped representation does facilitate the easy application of time-domain masking and fast transforms. Since the lack of a periodic structure means that time-synchrony is of lesser importance for the REW component than for the SEW, its preferred quantisation domain is the warped frequency domain, in which some perceptual properties may be exploited.

This work contributed to the following publications: [Chon00b], [Chon00c].

6.3.3 Performance at 4kbit/s

The 4kbit/s WMWI coder was compared to that of the standard coders G.729, a toll quality 8kbit/s CS-ACELP coder, FS-1016, a 4.8kbit/s CELP coder, and the 2.4kbit/s FS MELP coder using a Mean Opinion Score (MOS) test. While the coder did not achieve toll quality, it did show improved performance above FS-1016 and MELP, gaining a MOS score of 3.53.

6.4 Further Work

As the drive for a 4kbit/s toll-quality speech coding algorithm continues, a coder which combines the favourable attributes of a parametric coder and a waveform coder shows great potential to offer a solution. The principles of the WMWI coder proposed in this thesis form a strong basis for a successful 4kbit/s coding scheme.

There are several areas of the WMWI coder which could be further improved. Firstly, a main contributor to delay is the pitch track formation technique. This requires all pulses of the current frame, plus the first pulse of the future frame, to be located. The detection of the final pulse may introduce considerable delay if the pitch length is large. The technique could be adapted to reduce the number of component autocorrelation functions contributing to the pitch estimate if the required look-ahead exceeds one frame, to limit the total pitch delay to 25ms. Alternatively, the pitch for the final period could be predicted using previous pitch estimates.

The CW surface decomposition provides another opportunity for delay minimisation. FIR filters, commonly used to decompose the signal, exhibit poor separation of frequency subbands and incur high delays. To eliminate the delay, a predictive, evolutionary domain decomposition scheme could be applied. The technique would decompose the CW into a two or more components, based on the degree of prediction of the current CW from previous CWs. Past decomposed components could also be considered, resulting in a backward adaptive prediction mechanism. The current decomposition delay allows additional pitch information for Unpulsed-to-Pulsed frames to be transmitted during the previous frame. However, the transmitted pitch value could be adaptively selected to minimise the

reconstruction error and achieve a similar level of accuracy (with the same bit allocation) without relying on filtering delays.

It would also be desirable to quantify the rate of change in parameters to allow the most efficient waveform update rate to be used (directly affecting the bit rate), as was implicit in the wavelet decomposition. This may involve adaptively identifying significant features of the CW, e.g. the central pulse region, and creating a measure of deviation across these components. This could then be translated into the required information rate.

Alternatively, rather than performing a decomposition based on speech evolution characteristics, transforms, such as [Luka00],[Klei00], could be applied to gain coding efficiency.

There is considerable knowledge of the perceptual properties of the human ear, defined by the Bark Scale and critical band analysis. These relate to conventional frequencies. While some research has been conducted on the perception of voiced and unvoiced sounds, it would be desirable to form a detailed analysis of perceptual characteristics of evolution frequencies. Identification of the important aspects of particular evolution frequencies allows methods for improved quantisation of these signals to be established. This contrasts current WI efforts whereby all components evolving faster than 20Hz are considered of relative insignificance, often resulting in loss of naturalness, characterised by buzzy or noisy sounding speech.

The size of pitch sub-ranges of the SEW codebooks could be optimised. Constant sub-range sizes presently used, results in greater misalignment of the negative pulse peak for periods of shorter pitch, due to the warping factor, *L/pitch*. Quantisation of

an overall gain and the ratio of decomposed components may also provide better performance than separate gain quantisation for the decomposed components.

Also, the perception of time-synchrony is of interest. The good performance of WMWI may be attributed to both matching the waveform detail, as well as time-synchronous representation of all pitch periods in their entirety. Correct positioning of pitch pulses was found to give a more natural, vibrant sound, which maintains the sudden "attacks" of speech. This is opposed to the "flatter" sounding quality of time-asynchronous speech, in which energy bursts are dispersed over time. It may be that time-synchrony is really only of significance during transitional sections of speech to achieve accurate representation of sudden energy changes.

Several other minor improvements can be made to the WMWI coder. In particular, no post-filtering has been applied to emphasise the formant structure and enhance perceptual quality, nor any pre-processing operations are in place to improve the robustness of the coder to noisy environments. In addition, the algorithm has not been optimised in terms of computational complexity. It is believed that significant complexity reductions can be achieved, particularly in the pulse detection technique, without a degradation in quality,.

In conclusion, WMWI possesses perfect reconstruction and the facility to separate perceptually different speech characteristics to gain coding efficiency as its main advantages. It achieves accurate signal analysis and exploits the similarity between adjacent waveforms. With additional refinements, the WMWI framework, based on the attributes of a successful low rate parametric coder combined with waveform coding objectives, represents a potential candidate for 4kbit/s speech coding technology.

References

[Ahma98] S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 5, pp. 495-501, Sept. 1998.

[Arge96] F. Argenti, V. Cappellini, A. Sciorpes, A. N. Venetsanopoulos, "Design of IIR linear phase QMF banks based on complex allpass sections," *IEEE Trans. Signal Processing*, vol. 44, no. 5, pp. 1262-1267, May 1996.

[Atal71] B. S. Atal and S. L. Hannauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655, Aug. 1971.

[Atal79] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Speech, Signal Processing*, vol. 27, no. 3, pp. 247-254, 1979.

[Atal82] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 614-617, 1982.

[Atal84] B. S. Atal, M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates," *Proc. Int. Conf.on Communications*, pp. 1610-1613, 1984.

[Basu95] S. Basu, C-H. Chiang, H. M. Choi, "Wavelets and perfect reconstruction subband coding with causal stable IIR filters," *IEEE Trans. Circuits and Systems II*, vol. 42, no. 1, pp. 24-38, Jan. 1995.

[Beri99] F. Beritelli, "A modified CS-ACELP algorithm for variable-rate speech coding robust in noisy environments," *IEEE Signal Processing Letters*, vol. 6, no. 2, pp. 31-34, Feb. 1999.

[Burn93] I. S. Burnett and R. J. Holbeche. "A mixed prototype waveform / CELP coder for sub 3kb/s," *Proc. IEEE Int. Conf. Acoust., Speech, Signal*

Processing, vol. 2, pp. 175-178, 1993.

[Burn95] I. S. Burnett and G. J. Bradley, "New techniques for multi-prototype waveform coding at 2.84kb/s," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 261-264, 1995.

[Burn97] I. S. Burnett, D. H. Pham, "Multi-prototype waveform coding using frame-by-frame analysis-by-synthesis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1567-1570, 1997.

[Camp89] J. P. Campbell, V. C. Welch and T. E. Tremain, "An expandable error protected 4800 bps CELP coder (US Federal Standard 4800 bps voice coder)," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 735-738, 1989.

[Chen87] J. H. Chen and A. Gersho, "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2185-2188, 1987.

[Chen92] J. Chen, R. Cox, Y. Lin, N. Jayant and M. Melchner, "A low-delay CELP coder for the CCITT 16kb/s speech coding standard," *IEEE Journ. Selected Areas Commun.*, vol. 10, pp. 830-849, June 1992.

[Chon97] N. R. Chong, I. S. Burnett, J. F. Chicharo, M. M. Thomson, "The effects of noise on the Waveform Interpolation speech coder," *Proc. IEEE TENCON*, Brisbane, Australia, vol. 2, pp. 609-612, Dec. 1997.

[Chon98a] N. R. Chong, I. S. Burnett, J. F. Chicharo, M. M. Thomson, "Use of the pitch synchronous wavelet transform as a new decomposition method for WI," *Proc. IEEE Int Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 513-516, 1998.

[Chon98b] N. R. Chong, I. S. Burnett, J. F. Chicharo, "An improved decomposition method for WI using IIR filter banks," *Proc. Int. Conf. Spoken Lang. Processing*, 1998.

[Chon99a] N. R. Chong, I. S. Burnett, J. F. Chicharo, "Low delay multi-level decomposition and quantisation techniques for WI coding," *Proc. IEEE Int Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 241-244, 1999.

[Chon99b] N. R. Chong, I. S. Burnett, and J. F. Chicharo "Adapting waveform interpolation (with pitch-spaced subbands) to facilitate vector quantisation", *Proc. IEEE Workshop on Speech Coding*, pp. 96-98, 1999.

[Chon00a] N. R. Chong, I. S. Burnett, and J. F. Chicharo, "A new waveform interpolation coding scheme based on pitch synchronous wavelet transform decomposition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 345-348, May 2000.

[Chon00b] N. R. Chong and I. S. Burnett, "Improved signal analysis and time-synchronous reconstruction in Waveform Interpolation coding," *Proc. IEEE Speech Coding Workshop*, 2000.

[Chon00c] N. R. Chong and I. S. Burnett, "Method and apparatus for time-warping a digitised signal to have an approximately fixed period," US Patent CR1029AC, Feb. 2000.

[Chon00d] N. R. Chong, and I. S. Burnett, "Method and apparatus for encoding and reconstructing the pitch track of a digitised time-varying waveform," Australian Provisional Patent, Apr. 2000.

[Chon00e] N.R.Chong and I.S.Burnett, "Accurate, critically-sampled characteristic waveform surface construction for waveform interpolation decomposition," accepted for publication in *IEE Electronics Letters*.

[Clar95] J. Clark and C. Yallop, *An Introduction to Phonetics and Phonology*, 2nd ed., Blackwell, Oxford, UK, 1995.

[Comb99] P. Combescure *et al.*, "A 16, 24, 32 kbit/s wideband speech coded based on ATCELP," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 5-8, 1999.

[Cox88] R. V. Cox, S. L. Gay, Y. Shoham, S.R. Quackenbush, N. Seshadri, N. S. Jayant, "New directions in sub-band coding," *IEEE Journ. Selected Areas Commun.* , vol. 6, no. 2, pp. 391-409, Feb. 1988.

[Das96] A. Das, V. R. Rao, A. Gersho, "Variable-dimension vector quantisation," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 200-202, July 1996.

[Daub92] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, CBMS Lecture Series,

1992.

[Davi86] G. Davidson and A. Gersho, "Complexity reduction methods for vector excitation coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3055-3058, 1986.

[Ekud99] E. Ekudden, R. Hagen, I. Johansson, J. Svedberg, "The adaptive multi-rate speech coder," *Proc. IEEE Workshop on Speech Coding*, pp. 117-119, 1999.

[Erik99] T. Eriksson, W. B. Kleijn "On waveform interpolation coding with asymptotically perfect reconstruction", *Proc. IEEE Workshop on Speech Coding*, pp. 93-95, 1999.

[Erik99b] T. Eriksson and H-G. Kang, "Pitch quantisation in low bit-rate speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 489-492, 1999.

[Evan93] G. Evangelista, "Pitch-synchronous wavelet representations of speech and music signals", *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3313-3329, 1993

[Flan72] J. Flanagan, *Speech Analysis, Synthesis and Perception*, New York, Springer-Verlag, 1972

[Gers90] I. Gerson and M. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8kbit/s," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 461-464, 1990.

[Gers92] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston : Kluwer Academic Publishers, 1992.

[Gibs80] J. Gibson, "Adaptive prediction in speech differential encoding systems," *Proceedings of the IEEE*, vol. 68, pp 488-525, Apr. 1980.

[Gott99] O. Gottesman, A. Gersho, "Enhanced waveform interpolative coding at 4 kbps," *Proc. IEEE Workshop on Speech Coding*, pp. 90-92, 1999.

[Gott00] O. Gottesman, A. Gersho, "High quality enhanced wavefrom interpolative coding at 2.8 kbps," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, pp. 1363-1366, 2000.

[Gran91] W. Granzow, B. S. Atal, K. K. Paliwal, J. Schroeter, "Speech coding at 4kb/s and lower using single-pulse and stochastic models of LPC excitation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 217-220, 1991.

[Grif88] D. Griffin, J. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1223-1235, Aug. 1988.

[Ha99] N. K. Ha, "A fast search method of algebraic codebook by reorderding search sequence, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 21-24, 1999.

[Hage99] R. Hagen, E. Ekudden, "An 8kbit/s ACELP coder with improved background noise performance," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 25-28, 1999.

[Haag95] J. Haagen, W. B. Kleijn, "Waveform Interpolation," in *Modern Methods of Speech Processing*, edited by R. Ramachandran and R. Mammone, Kluwer Academic Publishers, 1995.

[Hard91] J. Hardwick and J. Lim, "The application of the IMBE speech coder to mobile communications," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 249-252, 1991.

[Hein99] S. Heinen, M. Adratm, O. Steil, P. Vary, W. Xu, "A 6.1 to 13.3kb/s variable rate CELP codec (VR-CELP) for AMR speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 9-12, 1999.

[Herl93] C. Herley and M. Vetterli, "Wavelets and recursive filter banks," *IEEE Trans. Signal Processing*, vol. 41, no. 8, pp. 2536-2556, Aug. 1993.

[Hess83] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Berlin; New York: Springer-Verlag, 1983.

[Jaya84] N. S. Jayant, and P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice Hall, 1984.

[Kang85] G. S. Kang and S. S. Everett, "Improvement of the excitation source in the narrow-band linear predictive vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 377-386, Apr. 1985.

[Kang99] H-G. Kang, D. Sen, "Phase Adjustment in Waveform Interpolation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 261-264, 1999.

[Kang99b] H-G. Kang, D. Sen, "Embedded WI coding between 2.0 and 4.8 kbit/s," *Proc. IEEE Workshop on Speech Coding*, pp. 87-89, 1999.

[Klei88] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Improved speech quality and efficient vector quantization in SELP," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 155-158, 1988.

[Klei90] W. B. Kleijn and D. J. Krasinski. "Fast methods for the CELP speech coding algorithm," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 38, pp. 1330-1342, Aug 1990.

[Klei92] W. B. Kleijn, R. P. Ramachandran and P. Kroon, "Generalised analysis-by-synthesis coding and its application to pitch prediction," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 337-340, 1992.

[Klei93] W. B. Kleijn, P. Kroon, L. Cellario, D. Sereno, "A 5.85 kb/s CELP algorithm for cellular applications," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 596-599, 1993.

[Klei93b] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 386-399, Oct 1993.

[Klei94] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding," *IEEE Signal Processing Letters*, vol. 1, no. 9, pp. 136-138, Sept. 1994.

[Klei95] W. B. Kleijn, J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal, Elsevier 1995.

[Klei96] W. B. Kleijn, Y. Shoham, D. Sen, R. Hagen, "A low-complexity waveform interpolation coder," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 212-215, 1996.

[Klei98] W. B. Kleijn, H. Yang, E. Deprettere, "Waveform interpolation coding with pitch-spaced subbands," *Proc. 5th Int Conf. Spoken Language.*

Processing, 1998.

[Klei00] W. B. Kleijn, "A frame interpretation of sinusoidal coding and waveform interpolation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, pp. 1475-1478, 2000.

[Kohl97] M. A. Kohler, "A comparison of the new 2400 bps MELP Federal Standard with other standard coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1587-1590, 1997.

[Kois98] K. Koishida, G. Hirabayashi, K. Tokuda, T. Kobayashi, "A wideband CELP speech coder at 16kbit/s based on mel-generalized ceptral analysis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 161-164, 1998.

[Kroo86] P. Kroon, E. F. Deprettere, R. J. Sluyter, "Regular pulse excitation: A novel approach to effective and efficient multi-pulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054-1063, Oct. 1986.

[Kroo90] P. Kroon, B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 661-664, 1990.

[Kubi93] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," *Proc. IEEE Workshop on Speech Coding for Telecommun.*, pp. 35-36. 1993.

[Lebl93] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kbit/s speech coding," *IEEE Trans. Speech, Audio Processing*, vol. 1, no. 4, pp. 373-385, Oct. 1993.

[McAu95] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K.K. Paliwal, Elsevier 1995.

[McCr95] A. V. McCree, T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 242-249, July 1995.

[McCr96] A. McCree, K. Truong, E. B. George, T. P. Barnwell III, V. Viswanathan, "A 2.4kbit/s MELP coder candidate for the new U.S. Federal Standard," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 200-203, 1993.

[McCr98] A. McCree, and J. C. De Martin, "A 1.7kb/s MELP coder with improved analysis and quantisation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 593-596, 1998.

[Makh75] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.

[Makh78] J. Makhoul, R. Viswanathan, R. Schwatrz and A. W. F.Huggins, "A mixed-source model for speech compression and synthesis," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1577-1581, Dec. 1978.

[Mall98] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[Malv92] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, Norwood, MA, 1992

[Mark76] J. D. Markel and A. H. Gray, Jr, *Linear Prediction of Speech*, Springer-Verlag, Berlin Heidelberg, Germany, 1976.

[Marq90] J. S. Marques, I. M. Trancoso, J. M. Tribolet, L. B. Almeida, "Improved pitch prediction with fractional delays in CELP coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 665-668, 1990

[Moor97] B. C. Moore, *An Introduction to the Psychology of Hearing*, 4th ed., Academic Press, 1997.

[Naye94] K. Nayebi, T. P. Barnwell, III, M. J. T. Smith, "Low delay FIR filter banks: Design and evaluation," *IEEE Trans. Signal Processing*, vol. 42, no.1, pp. 24-31, Jan. 1994.

[Ohmu93] H. Ohmuro, T. Moriya, K. Mano, S. Miki, "Coding of LSP parameters using interframe moving average prediction and multi-stage vector quantisation, "*Proc. IEEE Workshop on Speech Coding for Telecomm.*, pp. 63-64, 1993.

[Okud98] M. Okuda, M. Ikehara, S. Takahashi, "Design of biorthogonal filter banks composed of linear phase IIR filters," *Proc. IEEE Int. Conf. Acoust.,*

Speech, Signal Processing, vol. 3, pp. 1453-1456, 1998.

[Paks99] E. Paksoy *et al.*, "An adaptive multi-rate speech coder for digital cellular telephony," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 193-196, 1999.

[Pali93] K. K. Paliwal and B. S. Atal, "Efficient vector quantisation of LPC parameters at 24bits/frame," *IEEE Trans. Speech, Audio Processing*, vol. 1, no. 1, pp. 3-14, Jan. 1993.

[Pan98] J. Pan, T. R. Fischer, "Vector quantisation of speech line spectrum pair parameters and reflection coefficients," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 2, pp. 106-115, Mar. 1998.

[Prin95] J. Princen, "The design of non-uniform modulated filter banks," *IEEE Trans. Signal Processing*, vol. 43, no. 11, pp. 2250-2260, Nov. 1995

[Qian93] S. Qian, D. Chen, "Discrete gabor transform", *IEEE Trans. Signal Processing*, vol. 41, no. 7, pp. 2429-2438, July 1993.

[Rabi78] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, 1978.

[Rama87] R. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 7, pp. 937-946, 1987.

[Rama89] R. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 4, pp. 467-477, 1989.

[Rami99] M. A. Ramirez, M. Gerken, "A multistage search of algebraic CELP codebooks," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 17-20, 1999.

[Sala98] R. Salami et al., "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 2, pp. 116-130, Mar. 1998.

[Schn98] J. Schnitzler, "A 13.0kbit/s Wideband Speech Codec Based on SB-ACELP," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1,

pp. 157-160, 1998.

[Schr85] M. R. Schroeder and B. S. Atal, "Code-excited linear predictive (CELP): High quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937-940, 1985.

[Shlo98] E. Shlomot, V. Cuperman and A. Gersho, "Combined harmonic and waveform coding of speech at low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 585-588, 1998.

[Shoh91] Y. Shoham, "Constrained-excitation coding of speech at 4.8 kb/s," in *Advances in Speech Coding*, edited by B. S. Atal, V. Cuperman and A. Gersho, Kluwer Academic, Holland pp. 339-348, 1991.

[Sinh96] D. Sinha, J. D. Johnston, "Audio compression at low bit rates using a signal adaptive switched filterbank," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1053-1056, 1996.

[Skog97] J. Skoglund, W. B. Kleijn and P. Hedelin, "Audibility of Pitch-Synchronously Modulated Noise," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 51-52, 1997.

[Sohn99] J. Sohn and W. Sung, "A low resolution pulse position coding method for improved excitation modeling of speech transition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 265-268, 1999.

[Stac94] J. Stachurski and P. Kabal, "A pitch pulse evolution model for a dual excitation linear predictive speech coder," *Proc. 17th Biennial Symp. Commun.*, pp. 107-110, May 1994.

[Stac98] J. Stachurski, "A pitch pulse evolution model for linear predictive coding of speech," PhD Thesis, McGill University, Montreal, Canada, 1998.

[Stac00] J. Stachurski and A. McCree, "A 4kbit/s hybrid MELP/CELP coder with alignment phase encoding and zero-phase equalization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1379-1382, 2000.

[Stra96] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996

[Supp97] L. M. Supplee, R. P. Cohn, J. S. Collura, A. V. McCree, "MELP: The New Federal Standard at 2400 bps," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1591-1594, 1997.

[Tay97] D. Tay, "Design of causal, stable, IIR perfect reconstruction filter banks using transformations of variables," *Proc. IEEE Int. Symp. Circuits, Systems*, vol. 4, pp. 2425-2428, 1997.

[Thys97] J. Thyssen, W. B. Kleijn, and R. Hagen, "Using a perception-based frequency scale in Waveform Interpolation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1595-1598, 1997.

[Trem82] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology*, pp. 40-49, Apr. 1982.

[Trib79] J. Tribolet and R. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 5, pp. 512-530, Oct. 1979.

[Unse93] M. Unser, A. Aldroubi, M. Eden, "B-Spline signal processing: Part I – theory," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 821-848, Feb. 1993.

[Vaid90] P. P. Vaidyantathan "Multirate digital filters, filter banks, polyphase networks and applications: A tutorial," *Proceedings of the IEE*, vol. 78, no. 1, pp. 56-89, Jan. 1990

[Vaid93] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice-Hall Signal Processing Series, 1993.

[Vasi99] A. Vasilache, M. Vasilache, I. Tabus, "Predictive multiple-scale lattice VQ for LSF quantisation", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 656-660, 1999.

[Vett92] M. Vetterli, C. Herley, "Wavelets and filter banks: Theory and design," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2207-2232, Sept. 1992.

[Yagh97] K.Yaghmaie and A. M. Kondoz, "Multiband prototype waveform analysis synthesis for very low bit rate speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1571-1574, 1997.

[Yang98] H. Yang, W. B. Kleijn "Pitch-synchronous subband representation of the linear prediction residual of speech," *Proc. Int Conf. Acoust. Speech, Signal Processing*, 1998.

[Zeli79] R. Zelinski and P. Noll, "Approaches to adaptive transform coding at low bit rates," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, p. 89, 1979.